



Proceedings

Leonardo Bacelar Lima Santos and Marconi de Arruda Pereira

Dados Internacionais de Catalogação na Publicação

SI57a Simpósio Brasileiro de Geoinformática (11.: 2022: São José dos Campos, SP)

Anais do 23º. Simpósio Brasileiro de Geoinformática, São José dos Campos, SP, 28 a 30 de novembro de 2022. / editado por Leonardo Bacelar Lima Santos, (CEMADEN), Marconi de Arruda Pereira (UFSJ) – São José dos Campos, SP: MCTIC/INPE, 2022.

On-line

ISSN 2179-4847

1. Geoinformação. 2. Bancos de dados espaciais. 3. Análise Espacial. 4. Sistemas de Informação Geográfica (SIG). 5. Dados espaço-temporais. I. Santos, L.B.L II. Pereira, M. A. III. Rosim, S., IV. Título.

CDU:681.3.06

Preface

The papers presented at the XXIII Brazilian Symposium on Geoinformatics, GEOINFO 2022, are included in this volume of proceedings. The Federal University of Sao Joao Del-Rei (UFSJ), the National Institute for Space Research (INPE), and the National Center for Monitoring, Warning of Natural Disasters (CEMADEN) organized this edition. This is the first presencial edition of GEOINFO since the COVID-19 pandemic.

This issue includes 5 database descriptions, 5 tools, 31 short papers, 17 full papers, and 17 full papers. We appreciate the Program Committee's thorough review of the works submitted. The long papers were presented during oral sessions, while the short papers, databases and tools descriptions were presented in poster sessions. The databases and tools are available for download at the links in texts of their respective papers.

GEOINFO 2022 presented two guest lectures. Dr. Natalie Aubet from the Centro Sur Regional Technological Institute, Universidad Tecnologica del Uruguay gave a presentation on the first day titled *Groundwater Vulnerability Assessment Using GIS, EO Data: the Recharge Area of the Guarani Aquifer System in Uruguay*. The second day featured a presentation on *Science, Technology, and Ethics* by Dr. Renato Janine, president of the Brazilian Society for the Progress of Science.

The audience could also attend to two mini-courses: *Google Earth Engine* and *Brazil Data Cube - EO Data Cube Products, Time series analysis*. A special session on *Digital Technologies for the Environment: an SBC Manifesto* and a roundtable on *Postgraduate Programs with Lines of Research in Geoinformatics* were also held at GEOINFO 2022.

Our sincere regards,

Sergio Rosim, INPE
Leonardo Bacelar Lima Santos, CEMADEN
Marconi de Arruda Pereira, UFSJ
Chairs, GEOINFO 2022

Conference Committee

General Chair

Sergio Rosim, INPE

Program Chairs

Leonardo Bacelar Lima Santos, CEMADEN

Marconi de Arruda Pereira, UFSJ

Technical Sessions Chairs

Bruno Guazzelli Batista, UNIFEI

Clodoveu Augusto Davis Junior, UFMG

Flavia da Fonseca Feitosa, UFABC

Laercio Namikawa, INPE

Liana O. Anderson, CEMADEN

Marcos Aurélio Santos da Silva, Embrapa

Local Organization

Adriana Gonçalves, INPE

Gislaine Aparecida de Oliveira Faria, INPE

Organized by

INPE - National Institute for Space Research

CEMADEN - National Center for Monitoring, Warning of Natural Disasters

UFSJ - Federal University of São João Del-Rei

Supported by

CNPq - Conselho Nacional de Desenvolvimento Científico e Tecnológico

SELPER - Associação de Especialistas Latinoamericanos em Sensoriamento Remoto

Sponsored by

Axelspace

DTI Digital

Geopixel

Tagna

Visiona

Program Committee

Marconi A. Pereira (Chair) - USFSJ
Leonardo B. L. Santos (Chair) - CEMADEN
Alan J. Salomão Graça - UERJ
Alana Neves - INPE
Ana Cunha - CEMADEN
A. Miguel V. Monteiro - INPE
Alber Sánchez - INPE
Ana Clara M. Moura - UFMG
André Hirsch - UFSJ
Angelica Nardo - CEMADEN
Camilo D. Rennó - INPE
Carlos A. Felgueiras - INPE
Carolina M. D. Pinho - UFABC
Celso H. L. Silva Junior - INPE
Claudia Bauzer Medeiros - UNICAMP
Claudia P. Krueger - UFPR
Claudia R. Slutter - UFRS
Claudio Almeida - INPE
Claudio Campelo - UFCG
Claudio S. Baptista - UFCG
Clodoveu A. Davis-Jr - UFMG
Dalton M. Valeriano - INPE
Daniel Vila - INPE
Flavia F. Feitosa - UFABC
Francielle Cardozo - UFSJ
Gabriel Pereira - UFSJ
Gladston Moreira - UFOP
Gilberto R. Queiroz - INPE
Giovanni V. Comarela - UFES
Gustavo Marcatti - UFSJ
Haron A. M. Xaud - EMBRAPA
M. Isabel S. Escada - INPE
Jorge A. P. Campos - UNIFACS

Julio C. D. M. Esquerdo - EMBRAPA
Jugurta Lisboa Filho - UFV
Karine R. Ferreira - INPE
Karla Borges - PRODABEL
Klecia Massi - UNESP
Laercio Namikawa - INPE
Leonardo S. Bins - INPE
Liana O. Anderson - CEMADEN
Lincoln Alves - INPE
Lubia Vinhas - INPE
Luciana Londe - CEMADEN
Manoel A. Sousa Junior - UFSM
Marcelo Zeri - CEMADEN
Marco A. Casanova - PUC-Rio
Marcos Adami - INPE
Michel E. D. Chaves - INPE
Michelle Picoli - U. Catholique de Louvain, Belgium
Paula Debiasi - UFRJ
Pedro R. Andrade - INPE
Rafael C. Santos - INPE
Raul Q. Feitosa - PUC-Rio
Renato Fileto - UFSC
Rogerio Negri - UNESP
Sergio D. Faria - UFMG
Sergio Rosim - INPE
Sidnei J. S. Sant'Anna - INPE
Silvana Amaral - INPE
Sonaira Silva - UFAC
Thales S. Korting - INPE
Tatiana M. Kuplich - INPE
Tiago G. S. Carneiro - UFOP
Vander Freitas - UFOP



Universidade Federal
de São João del-Rei



ASSOCIAÇÃO DE ESPECIALISTAS
LATINOAMERICANOS EM
SENSORIAMENTO REMOTO
SELPER - BRASIL

dti

tagna



VISIONA
Tecnologia Espacial

AXELSPACE

Contents

Full Papers	1
Comparison of regional scale burned area products for southwestern Brazilian Amazonia	1
Generating POI type embeddings based on variable Word2Vec sentences	15
Feature extraction of spatial panel data with autoencoders for clustering the Brazilian agricultural diversity	27
A new big data architecture for efficient processing of spatiotemporal data using machine learning	39
A survey on persistence strategies for semantically Enriched Trajectories	51
Exploring the diversity of informal settlements in Brazil: a clustering approach to qualitative field survey data	63
Application of GEOBIA for urban area classification using optical images from the PE3D project	74
A tool to predict the growth of urban regions based on QGIS/MOLUSCE using MapBiomass image time series	86
Remote sensing, machine learning on anomaly detection at high spectral, temporal dynamics regions in Brazil	99
Characterization of fracture geometry with computed microtomography rock images	111
Fire patterns in the Pantanal biome	123
Database of 2019 Brazilian oil disaster: an overview of a dataset, its application on an Artificial Intelligence monitoring system	133
Georeferenced multipurpose real estate re-registration support system	144
A QGIS plugin for BONDS project: integrating field data with geographical, remote sensing, health information	156
Spatiotemporal Representation of the territorial evolution of Brazil	168
Territorial relations between deforestation, Rural Environmental Registry (CAR) in the Amazon biome using free software QGIS/PostgreSQL/PostGIS, Data Warehouse Structure	180
Deep learning automated workflow for cloud segmentation in remote sensing images	192
Short Papers	204
Mapeamento de corte seletivo na Amazônia Brasileira usando imagens Planet de alta resolução e inteligência artificial: estudo de caso na Floresta Nacional do Jamari	204

Um algoritmo de máxima concordância para harmonizar legendas de mapas de uso e cobertura da terra	211
Acessibilidade a escolas na região do Grande ABC: desigualdades socioespaciais	217
Statistical analysis of a spatially explicit hydrological forecasting model using remote sensing of the atmosphere	223
Deforestation increases within full protected areas of the Brazilian Amazon	229
Análise da cobertura da terra no Parque Nacional Yasuní, Equador, usando Random Forest e computação em nuvem	235
Vazios de informação & informalidade: web-scraping para identificação de assentamentos precários	241
Integrando métodos de análise durante a coleta de amostras de Uso e Cobertura da Terra	247
Mapeamento colaborativo do nível de ruído em Guayaquil durante o COVID-19	253
Previsão da probabilidade de fogo: ocorrências pós-alerta na Amazônia Maranhense	259
Uso de cubos de dados de observação da Terra no mapeamento de áreas queimadas	264
Mapeamento da susceptibilidade à incêndios através de Sensoriamento Remoto no município de São Félix do Xingu, Pará	270
Ionospheric scintillation awareness for GPS, other constellations by probability maps	275
Serviço Web Para extração de métricas fenológicas para agricultura usando séries temporais de imagem de satélite	281
Modeling, validation of elevation data obtained by GNSS, drone devices for managing floodings in coastal lands caused by climate change	287
Pacote ‘Contraste’: aplicação de técnicas de contraste baseado no uso e cobertura da terra	293
Modeling the dynamics of moving regions for knowledge discovery	299
Agrupamento de endereços georreferenciados para a delimitação de áreas postais	305
Contribuição de produtos de área queimada na Amazônia Maranhense: proposta de avaliações combinadas	311
The effects of COVID-19 on Chinese commuting patterns in early 2020	317
Uso do Google Earth Engine como instrumento de identificação de anomalias climáticas no Brasil	323
Desastres ambientais utilizando dados públicos	329
Analysis of the watersheds of National Interconnected System (SIN) using Complex Network	335
WebODM para o processamento de imagens aéreas obtidas por drone em uma área rural: como alternativa aos softwares proprietários	341
FMASK algorithm applied to Landsat images: a study case in the region of Santarém, Pará, Brazil	347
Controle da qualidade cartográfica de basemaps: avaliação da Completude	353

Analysis of mobility network metrics in the spread of COVID-19 in the Metropolitan Region of Vale do Paraíba, Litoral Norte (RMVPLN)	360
Um método de verificação de conformidade do plantio realizado no contexto do seguro ural no Brasil com Zoneamento Agrícola de Risco Climático (ZARC)	366
Padrões espaço-temporais da ocorrência do fogo na Amazônia brasileira	372
Rain-related Tweets in flooding days: a brief analysis on Tamanduateí river basin	378
A convolutional LSTM neural network for precipitation nowcasting based on weather radar data	384
Demonstrations	389
Daily Weather pattern for São Paulo city: from 1st January to 31st March, 2019	389
Development of a data set for distributed hydrological simulation of Paraíba do Sul Basin	395
Mobility database of the Metropolitan Region of Belo Horizonte, Brazil	401
An urban flood database for Disaster Risk Reduction studies in São Paulo, Brazil	406
Towards generating a traffic slowness geospatial dataset of São Paulo city	411
Q-CleanCAR: Um complemento QGIS para remoção guiada de sobreposições de polígonos de imóveis rurais do Cadastro Ambiental Rural (CAR)	417
Innovative tools for TerraHidro: an intuitive GIS user interface	423
Implementação de ferramentas em software livre para avaliação da acurácia posicional em produtos cartográficos	429
Sistema de geração de dados faltantes em séries temporais climatológicas usando triangulação, aprendizado e meta-aprendizado de máquinas	435
OpenStreetMap tools for humanitarian aid	441
Index of authors	447

Comparison of regional scale burned area products for southwestern Brazilian Amazonia

Débora Joana Dutra^{1*}, Liana Oighenstein Anderson², Philip Martin Fearnside³,
Aurora Miho Yanai⁴, Paulo Maurício Lima de Alencastro Graça⁵, Ricardo
Dalagnol⁶, Ana Carolina Moreira Pessoa⁷ and Luiz Eduardo Oliveira e Cruz de
Aragão⁸

^{1,2,7} National Center for Monitoring and Early Warning of Natural - 12247-016 - São
José dos Campos – SP - Brazil

^{3,4,5} National Institute for Research in Amazonia - 69067-375 - Manaus- AM - Brazil

⁶ NASA-Jet Propulsion Laboratory and University of California - CA 91109 – Oak
Grove Dr - Pasadena - USA

⁸ National Institute for Space Research, 12227900 - São José dos Campos - SP

{ddutra.ambiental@gmail.com; liana.anderson@cemaden.gov.br;
philip.fearnside@gmail.com; yanai@inpa.gov.br;
pmalencastro@gmail.com; ricds@hotmail.com; acmoreirapessoa@gmail.com
and luiz.aragao@inpe.br}

Abstract. Fires affect the Amazon Forest and cause a variety of socioenvironmental problems. Monitoring the dynamic of forest fires is important to support both combatting and preventing wildfires. However, many studies reported discrepancies in the quantification of forests affected by fires, especially in the tropics. In this study, we aimed at assessing four operational burned area products. We defined a study area located in the southwest Brazilian Amazon and selected the year 2019. The assessment focused on evaluating the product's relative performance, stratified by forest and non-forest areas. The four products showed divergence, by up to 90.61% in their estimates of the total burned area (MAPBIOMAS with more extensive burned area mapping and MCD64A1 with more minor burned area mapping).

Key-words: forest fire, burned area, land use cover change, geospatial analyses

1. Introduction

The human species is the main agent that is transforming natural ecosystems and causing profound changes in the landscape [Berlinck and Batista 2020], especially when associated with fire. The impact of fire in Brazil has increased, especially in recent years [Alencar et al. 2022]. Understory forest fires make the forest more susceptible to future fire events, especially in the Amazon rainforest region where this type of event is naturally rare [Bush et al. 2008]. This scenario can have various consequences for the ecosystem, such as impacts on biodiversity [Mataveli et al 2017, 2021 a,b], economic losses [Mendonça et al. 2004] and climate change [Aragão et al. 2018; Silva Junior et al. 2019; Carvalho et al. 2021; Aragão, Silva-Junior, and Anderson 2020].

Amazon rainforest is an important global climate regulator, acting as a provider of environmental services [Fearnside 2008], including the forest's role as a carbon stock and as a regulator of rainfall in South America [Leite-Filho et al. 2021]. This forest has been faced with deforestation and forest degradation propagated by the fire of anthropic

origin, which causes human health problems [Campanharo et al. 2022] and imbalances in the hydrological and carbon cycles [Maraseni et al. 2016; Prentice et al. 2011; Leite-Filho et al. 2021; Shakesby and Doerr 2006]. Therefore, it is important to develop actions to monitor fire-related activities to support measures for preventing and mitigating environmental impacts in the Amazon rainforest [Mataveli et al. 2021 a,b; Andrade et al. 2020].

The use of remote sensing techniques has allowed the development of methodological approaches to detect and monitor burned areas, especially in forest regions [Anderson et al. 2015; Giglio et al. 2018; Shimabukuro et al. 2015; Penha et al. 2020]. These methodologies have several specificities that can achieve different purposes and have different scales and spatial resolutions, leading to different results for the spatial distribution, time, size, and frequency of burned area [Mouillot et al. 2014; Long et al. 2019]. Diversity in mapping creates the need to use a comparison tool among the burned area products because it allows evaluation of the mapping according to performance [Humber et al. 2019; Padilla et al. 2015], especially when there are no field validation points in the region. In this aspect, it is important to understand that comparisons between products have a variety of limitations [Pessôa et al. 2020] and it is necessary to assume that all methodologies are only providing an approximation of the real conditions [Humber et al. 2019]. These comparisons should be used as a complement to the product validation process [Pessôa et al. 2020]. Due to these limitations, the comparison analysis considers the disadvantages and advantages of the products for different purposes [Pessôa et al. 2020; Anderson et al. 2015]. The analyses must be carried out according to the product specifications, and this information must be balanced in choosing the data to be included in the analysis and the final objective of the data use [Pessôa et al. 2020; Long et al. 2019; Boschetti et al. 2020].

Although several studies have assessed the differences in the results of burned-area products [Penha et al. 2020; Mataveli, Chaves, et al. 2021; Humber et al. 2019; Anderson et al. 2015; Pessôa et al. 2020], their inter-comparison suggests a spatially heterogeneous performance, which means that a prior evaluation of a site should be carried out before selecting a product. Here we assessed four operational burned-area products [MAPBIOMAS, MCD64A1, GABAM, and GWIS] for the southwest Brazilian Amazon – a region increasingly under threat of fires. The specific objectives were: (i) to evaluate the similarities and differences among operational burned-area products in forest and non-forest areas in the municipality of Boca do Acre, state of Amazonas, and (ii) to analyze the spatial similarities and differences between the products.

2. Materials and Methods

2.1 Study area

The study area focus on Boca do Acre Municipality, the State of Amazonas, including a buffer of 25 km around the, covering a total area of 40,777 km². The buffer area covers parts of the municipalites of Paiuni (19.42%), Lábrea (5.42%), Acrelândia (1.91%), Senador Guimard (16.16%), Porto Acre (79.01%), Bujari (28.18%), Sena Madureira (9.58%) and Manoel Urbano (13.68%). The area encompasses portions of seven indigenous territories: Camicua, Igarapé Capana, Inauini/Teuini, Boca do Acre, Apurinã, Peneri/Tacaquiri, and Seruini, Mariene, and five conservation units (protected areas for biodiversity): Mapiá-Inauini National Forest, Purus National Forest, and the Arapixi Extractive Reserve. According to PRODES data for 2019 [Assis et al. 2019], the regions contain 33,335.80 km² of forest (Figure 1), characterized by dense rainforest

(Amazon Forest), mosaics of oligotrophic woody vegetation, and ecotone areas [Barni et al. 2015] with Köppen classification system of Af (equatorial forest climate) [Alvares et al. 2013].

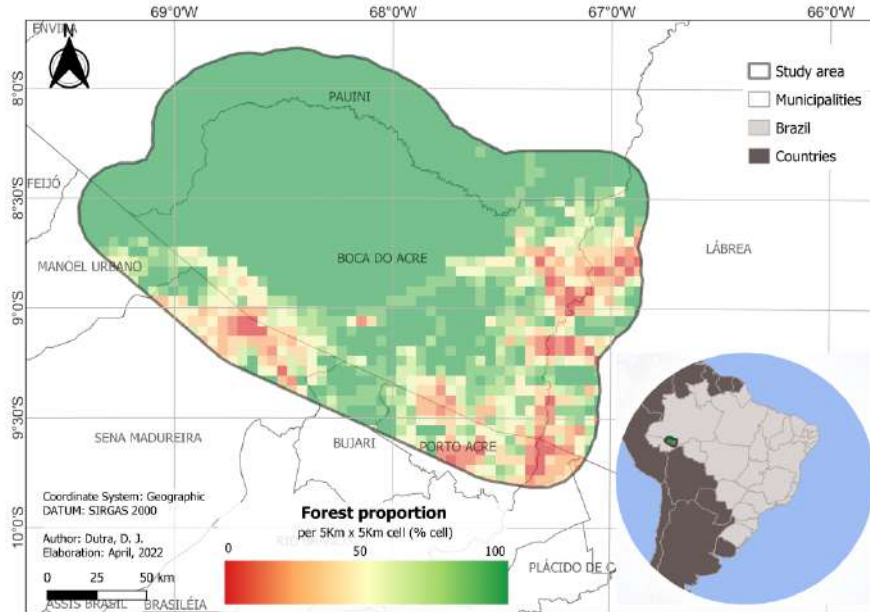


Figure 1 - Study area located in the Boca do Acre, municipality, state of Amazonas. Forest proportion in a 5x5 km grid cell, extracted by the Amazon Forest Deforestation Calculation Program (PRODES) forest mask of 2019 used to select burned areas

2.2 Data

We considered three global burned area products (MCD64A1 [Giglio et al. 2018], GABAM [Long et al. 2019], and GWIS [Boschetti et al. 2020]) and one national product MAPBIOMAS [Arruda et al. 2021]) for the comparative evaluation of burned area detection (Table 1). The choice of products took into account the spatial scale to compare products and analyze the influence of increased resolution in burned-area detection. For this, we used one low-resolution product that has been widely used in the literature (MCD64A1), one product in vectorial format (GWIS) and two products, with higher spatial resolution (30 m: MAPBIOMAS and GABAM). The year 2019 was selected for this analysis.

Table 1 - Specifications of the burned area products.

Name	Developer	Scale	Temporal Scale	Sensor/Data	Spatial resolution	Reference
GABAM	Institute of Remote Sensing and Digital Earth—Chinese Academy of Sciences	Global	1985-2020	Landsat series	30m	(Long et al. 2019)
GWIS	Group on Earth Observations (GEO) and Copernicus Work Programs	Global	2001-2020	MCD64A1, MODIS, Copernicus-Proba-V and Fire CC1	Vector data	(Boschetti et al. 2020)
MAPBIOMAS	MAPBIOMAS	National (Brazil)	2000-2020	Landsat series	30m	(Alencar et al. 2022; Arruda et al. 2021)
MCD64A1	NASA	Global	2000-present	MODIS (surface reflectance and active fires)	500m	(Giglio et al. 2018)

2.3 Analysis

We calculated the total burned area estimated by each product in the vectorial analyses, with separation by class in the whole study region. For this, the PRODES land-cover data from 2019 [Assis et al. 2019] were used to separate the landscape class into forest and non-forest. In the process, we used the SQL command and PyQGIS (gdal library) in QGIS 3.22.6 software (Qgis 2019) and the 'rgeos' package [Bivand and Rundel 2018] in RStudio statistical software [R Core Team 2021]. Furthermore, we created a regular grid with a spatial resolution of 25 km² (5 km × 5 km) for the matrix analysis of burned-area products. In the process, we considered the proportion of the burned area detected by each product per grid cell. The grid tools and SQL commands in QGIS software were used to process the data in the system.

Regarding the statistical analysis, we executed the non-parametric Kolmogorov–Smirnov two-sample test [Smirnov 1939] to compare the six possible combinations between the burned-area products. For this, we used the 'raster' package [Hijmans 2017] of RStudio software [R Core Team 2021]. We used conditional, repeating structures in a bootstrap approach to creating 10,000 interactions of 10% of the total cells that were randomly selected with replacement in each execution of the conditional structure. In this process, we considered only cells that presented burning detection by at least one. Finally, we considered the mean and standard deviation of the 10,000 p-values resulting from the interactions. For the spatial analyses, we converted the regular grid to raster format with information on each burned area, including a two-by-two statistical comparison using the fuzzy numerical method implemented in the “calc reciprocal similarity map” functor [Dinamica EGO Team 2020] of DINAMICA EGO 6 software [Leite-Filho et al. 2020]. The fuzzy method analyses used the spatial similarity between pairs of cells in two numerical maps, using the neighborhood (window size of 3 lines by 3 columns) to calculate the similarity of each cell [Dinamica EGO Team 2020]. Furthermore, the value interval of results is between 0 (fully distinct) and 1 (fully identical).

3. Results

We detected a total of 7167 km² burned area mapped in 2019 by the products in the study area, of which 264 km² (3.69%) was in the forest area and 6903 km² (96.31%) was in the non-forest area (Figure 2). In forest areas, GWIS and MCD64A1 detected the quantified larger extent: 90 km² (33.93%) and 100 km² (37.95%), respectively. In the non-forest areas, GABAM and MAPBIOMAS detected a larger extent of 3111.18 km² (45.07%) and 3338.98 km² (48.37%), respectively.

The burned area (Figure 3) showed the greatest difference occurring between MAPBIOMAS and MCD64A1, where MCD64A1 mapped 90.61% (3062 km²) less burned area in comparison with MAPBIOMAS. Compared to MCD64A1, MAPBIOMAS detected a 30.43% larger area (3122.33 km²) of burned in the non-forest area and a 30.43% smaller burned area (60.33 km²) in the forest area. Two of the products (GWIS/MCD64A1 and GABAM/MAPBIOMAS) produced very similar results. MCD64A1 detected only 2.46% (8 km²) less total burned than GWIS, but detected more burned area in the forest class (4.02% or 100.2 km²) and less burned area in the non-forest class (4.02% or 29.1 km²). MAPBIOMAS detected 6.93% (234 km²) more total burned area than GABAM, which detected 8% (228.72 km²) in non-forest areas and 8% (5.27 km²) in forest areas. The GWIS and MCD64A1 showed non-significant differences at a 95% confidence level ($p > 0.05$), Table 2. In this process, the bootstrap approach resulted in 99,37% of 10000 interactions was non-significant ($p >$

0.05). Other combinations among the burned area products present 100% significant p-values at a 95% confidence level ($p < 0.05$).

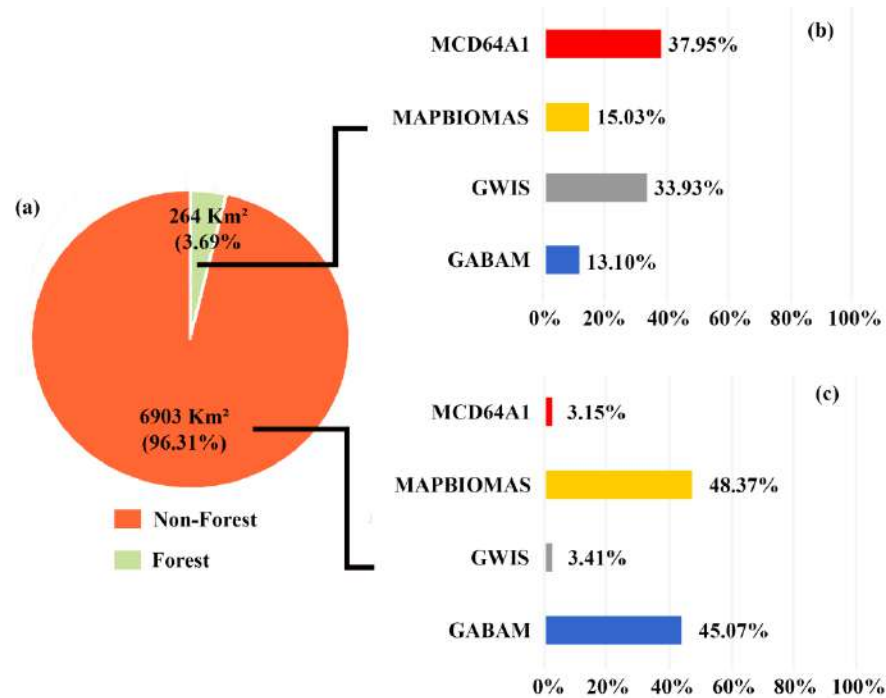


Figure 2 - (a) percentage of the total burned area detected by any of the products that are in forest and non-forest. (b) Percentage of the area classified as burned by each product in the forest area, and (c) Percentage of the area classified as burned by each product in the non-forest area.

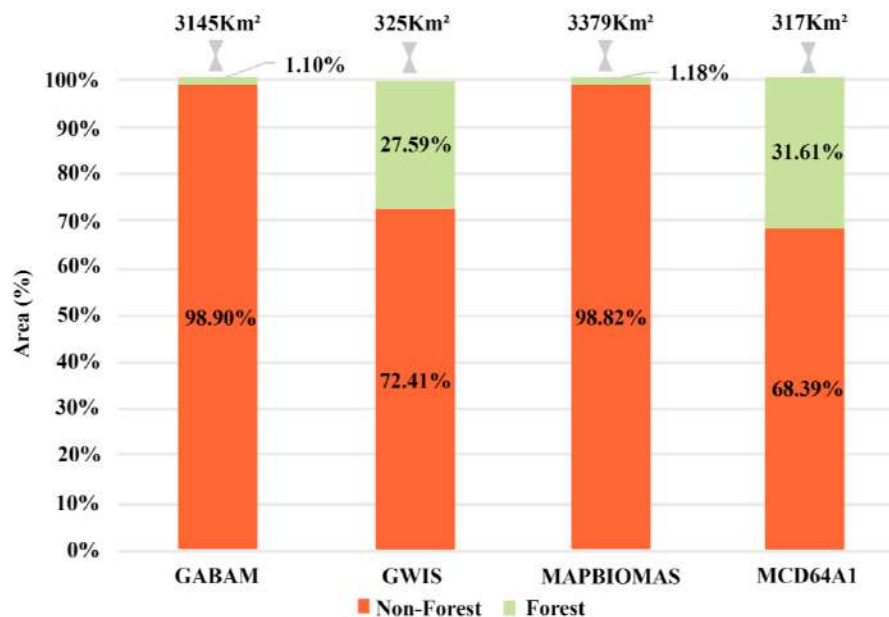


Figure 3- Total burned area mapped by GABAM, GWIS, MAPBIOMAS, and MCD64A1 and percentage burned area in the forest and non-forest classes.

Table 2 - Mean and standard deviation of p-values resulted from 10,000 iterations of Kolmogorov- Smirnov two-sample tests, raffling different samples of 10% of the total grid cells of 25 km².

	GABAM X GWIS	GABAM X MAPBIOMAS	GABAM X MCD64A1	GWIS X MAPBIOMAS	GWIS X MCD64A1	MAPBIOMAS X MCD64A1
Mean	6.97E-05	3.22E-05	1.04E-04	2.40E-08	5.94E-01	3.13E-08
Sd	2.96E-04	6.24E-05	3.18E-04	5.27E-07	2.91E-01	1.67E-07

Regarding the spatial analysis, the four products showed divergence, mainly in the northeast region of the study area that contains the largest presence of forest (Figure 4). Among the four products, MAPBIOMAS mapped the greatest amount with burned-cell proportions between 20% and 50% in grid cells (Figure 5). Compared with GABAM, with the same spatial resolution, the MAPBIOMAS project identified a greater number of burned cells in the range from 1.475km² (59 grid cells) to 110.975km² (4439 grid cells).

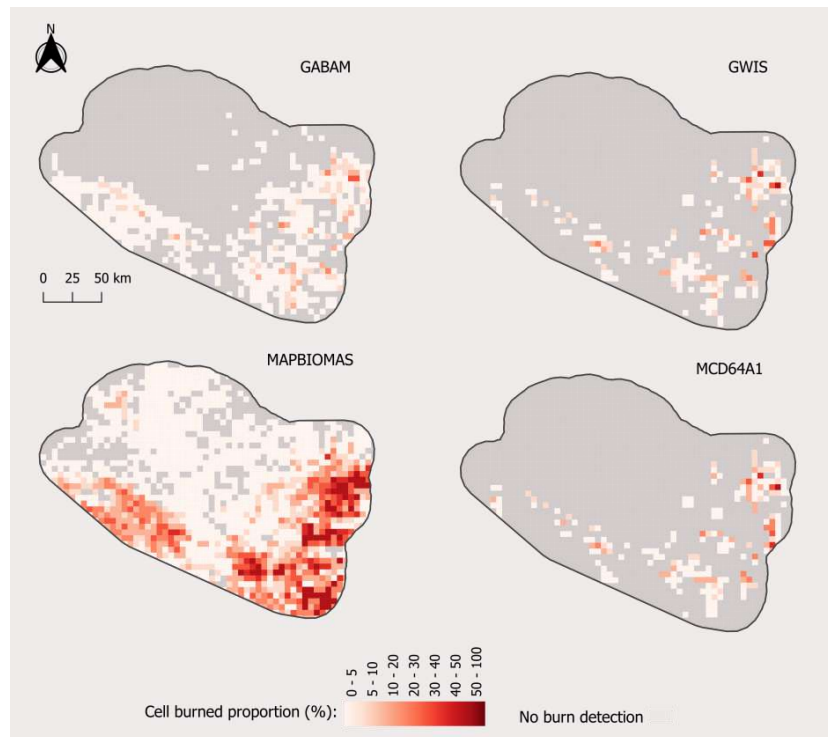


Figure 4 - Burned area spatialization in a 5 km × 5 km regular grid. Each grid cell contains the burned proportion indicated by the color gradient.

Number of cells in different burned area proportion class in a grid of 5km x 5km							
	0 - 5	5 to 10	10 to 20	20 to 30	30 to 40	40 to 50	>50
GABAM	443	57	22	3	2	0	0
GWIS	137	42	21	6	5	1	1
MAPBIOMAS	4882	2075	2351	1619	1223	724	59
MCD64A1	128	32	25	8	3	2	1
Total	5590	2206	2419	1636	1233	727	61

Figure 5 - Number of cells in different burned proportion class

We identified that the similarity index was medium to high (≥ 0.8) between the burned area products (Figure 6). Considering the whole study area, the results present values from 0.4 to 0.95. In the results, we observed a pattern of lower values of the similarity index when MAPBIOMAS is considered in comparison with the other products. These results can be explained by MAPBIOMAS presenting the largest extent of mapped burned area, which allows identifying burned areas that other products did not map. The other products presented higher similarity due to the reduced extent of the burned area mapped, which makes them more likely to be similar during the analysis.

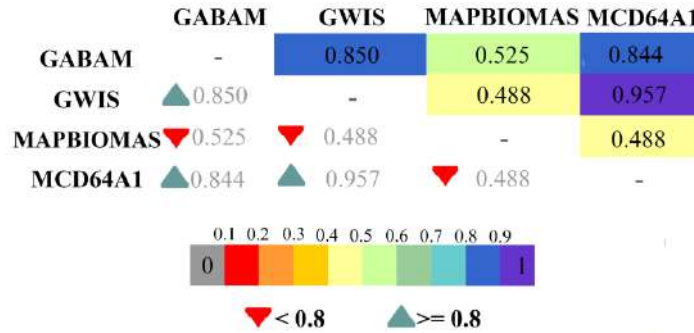


Figure 6 - Overall similarity for each burned area product comparison pair, considering the whole area

Spatially, the similarity varied with values between 0 and 1 (Figure 7). This analysis allows the identification of more cohesive areas, or not, since the similarity indices only registered the general average of the region. We identified the lowest values when MAPBIOMAS is considered in the analysis. In this process, the southeastern portion of the study area showed lower values between 0.7 and 0.9, and the northwest present higher values in this interval. The opposite relationship occurs with the other products, which demonstrate higher values in the southeastern portion of the study area.

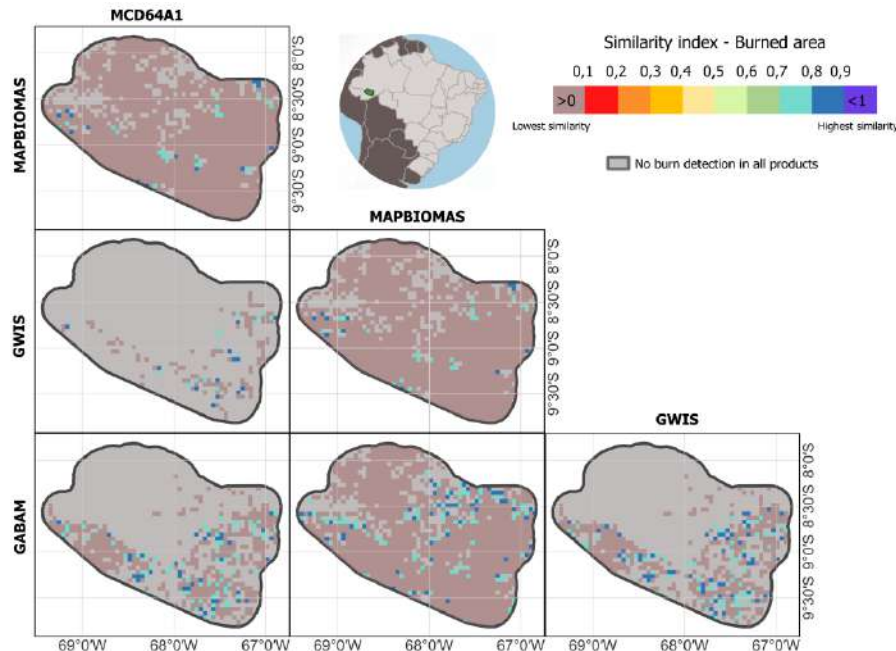


Figure 7 - Similarity maps for each burned area product comparison pair

We observed that MAPBIOMAS registered a low similarity index, mainly in the northwestern portion of the study area, where the product showed the best performance compared to the other products, and identified burned areas in the eastern and southeastern portions of the study area. In this analysis, we found that MCD64A1 and GWIS showed the burned area in the same portions of the study area as the other products, mainly in the eastern portion (Figure 8b). In contrast to GWIS and MCD64A1, GAMBAM presents many burned areas in the same portions of the study area as MAPBIOMAS, especially in the southwestern portion (Figure 8a). Regarding the scale resolution, for the GWIS and MCD64A1 products (lower resolution), the detected areas were mainly in patches in the eastern and southeastern portions of the study area near the edges of the study area. The high-resolution products (GABAM and MAPBIOMAS) identified more burned areas, mainly in the forest region (located in the northeastern portion of the study area), and presented the same burned areas as the other products in the eastern and southwestern portions of the study area.

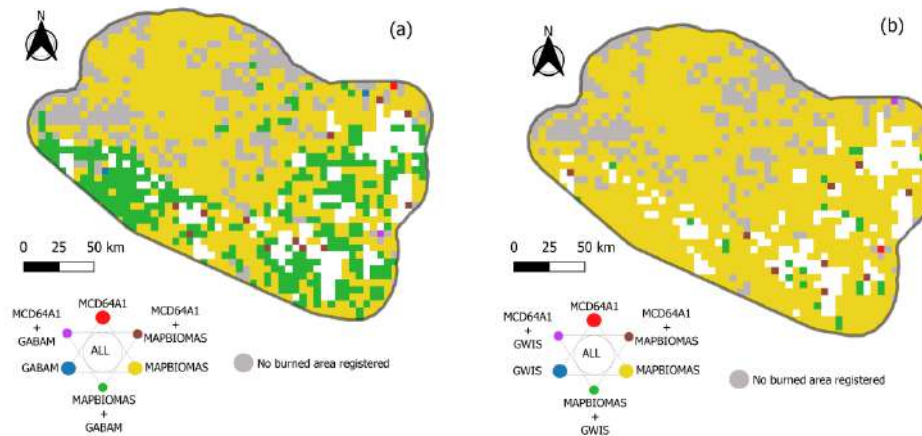


Figure 8 - Confusion maps considering (a) the GABAM, MCD64A1, and MPBIOMAS burned-area products, and (b) the GWIS, MCD64A1, and MPBIOMAS burned-area products

4. Discussions

Our results showed the importance of users understanding the characteristics of the burned area to choose the most appropriate product for their analyses, since each burned-area product may present a significant impact on the final result in studies at different scales, especially on regional scales. This is due to the characteristics and specifications that affect performances regionally, such as daily temporal resolution, spatial resolution, climatic conditions, and the algorithms themselves [Pessôa et al. 2020].

Regarding the mapping of burned areas, we identified two similar groups: GWIS/MCD64A1 and GABAM/MAPBIOMAS. Although the MCD64A1 product showed 2.46% to 90.61% less total burned area compared to the other products we analyzed, this product detected the most burned forest: 10.51% to 65.47% more than other products. Although some studies have shown that MODIS data can underestimate burned area by approximately 25% as compared to Landsat data [Morton et al. 2011; Roy and Boschetti 2009; Pessôa et al. 2020], we identified, on a municipality scale, that MAPBIOMAS and GABAM (Landsat data) generally underestimated burned scars,

especially in forest areas, when compared to MCD64A1 and GWIS (MODIS data). However, underestimation of burned area in products that use MODIS data as a reference was identified in non-forest regions, since MAPBIOMAS and GABAM data registered an increase of 92,95% to 93,50% in the burned-area mapping compared to MCD64A1 and GWIS. These differences can be associated with the spatial resolution of the data sources [Long et al. 2019].

We found that the characteristics of mapping can be delimited by different quantities of the burned area detected at regional scales. The high temporal resolution of MODIS data (used for MCD64A1 and GWIS), allows for greater data acquisition and less interference from clouds [Alonso-Canas and Chuvieco 2015; Pessôa et al. 2020], allowing the burned area date to be identified [Bush et al. 2008]. Thus, the frequency of MODIS data identifies the time elapsed since burning and the speed at which vegetation regenerates after the fire, these being important factors for monitoring tropical regions because the higher temporal frequency minimizes the effect of cloud cover and climatic conditions. This product has therefore been widely used in burned-area detection across the globe [Giglio et al. 2018; Justice et al. 2002]. The temporal resolution of Landsat data (used by GABAM and MAPBIOMAS) is 16 days, but the spatial resolution in the optical spectrum is higher (30 m). The higher spatial resolution from Landsat allows an improved definition of the boundaries of burned areas because these features avoid a mixture of burned and unburned patches in the same pixel [Long et al. 2019; Arruda et al. 2021].

We identified that this process can be overestimated when using national-scale products (MAPBIOMAS). The results demonstrate that MAPBIOMAS registered 11 times more burned area with small proportions when compared to a global product with the same resolution (25-km² grid cells). According to the developers of GABAM [Long et al. 2019], the overestimated values in the use of Landsat data can be associated with temporal resolution and cloud contamination [Long et al. 2019], which can become a limitation in tropical regions, where cloud cover is persistent and the vegetation recovery is quick. Regarding the MAPBIOMAS data, the developers recommend making some adjustments to the algorithm before applying multitemporal analysis in regions other than the Cerrado, which demonstrates that these data still need to be studied in tropical regions [Arruda et al. 2021].

Overestimation in MAPBIOMAS data may be associated with the application of the Deep Neural Network (DNN) methodology, which, according to the density of training samples per Landsat WRS-2 path/row map created by the product's developers, the study region is located in a scene where there was an intermediate separation of samples [Alencar et al. 2022]. Since the DNN uses pattern recognition to execute the algorithm [Safi and Bouroumi 2013; Langford, Kumar and Hoffman 2019], the low number of samples from the region may have helped in the overestimated result and reinforces the issues that in tropical regions adjustments are necessary for the product algorithm [Arruda et al. 2021]. In addition, we found that comparisons between MCD64A1 and GABAM data showed differences at a municipality scale when compared with the results of analyzes at a regional scale by the developers of MAPBIOMAS [Alencar et al. 2022]. Therefore, the use of images with lower resolution for mapping burned areas can be useful because these products have a higher temporal frequency and less influence on cloud cover [Giglio et al. 2018].

Regarding, the lower resolution products, such as MCD64A1 and GWIS, studies demonstrate that they are unable to adequately detect small fires (< 100 ha) [Rodrigues et al. 2019], which can cause the burned area to be underestimated [Giglio et al. 2018;

Justice et al. 2002], as we found in our results. Despite MCD64A1 presenting a significantly better detection of small burns (< 100 ha) than older products [Pessôa et al. 2020; Justice et al. 2002; Chuvieco et al. 2018], we found the underestimates of the burned area to be from 89.9% to 90.61% as compared to the high-resolution products (30 m). Therefore, GABAM and MAPBIOMAS detect small burned areas better, showing small burn proportions (associated with small burned patches) in the grid with 5-km × 5-km cells.

We observed different patterns when compared with similarities that can cross-validate each other through spatial analysis. We also identified the potential of the MAPBIOMAS product for mapping burned areas, although, for tropical regions such as the Amazon rainforest, future adjustments will be required, as described by the developers [Arruda et al. 2021]. Thus, in the absence of a national product that doesn't require adjustment in the algorithm, the global products still prove to be reliable for operationalization and analysis of socio-environmental loss related to tropical forest fire [Barlow et al. 2020].

5. Conclusions

A comparison of burned-area products allowed us to analyze the influence of spatial resolution in burned-area analysis at the regional scale. Accounting for the magnitude of difference, GWIS and MCD64A1 are the most similar products because they identified a smaller difference in the burned area compared to other products. The products that stand out the most are MAPBIOMAS and MCD64A1 due to a difference of 90.62% between the burned-area mappings. Regarding land use, we observed that the products with higher resolution (GABAM and MAPBIOMAS) showed smaller differences in burned area mapping than the products with lower resolution (GWIS and MCD64A1). This difference can be observed in products with the same origin (Landsat 8), where the MAPBIOMAS identified a greater amount of burned area than GABAM, and registered more small burned polygons. Despite the greater area mapped mapping of the burned area, MAPBIOMAS may have registered a greater interference from noise and the contribution of small polygons. Reference data, such as ground truthing or higher resolution in images would be necessary to further evaluate this. Thus, between comparisons at a municipality scale, the data from GABAM, GWIS, and MCD64A1 were the most similar for mapping the burned area in the study region, despite differences in spatial resolution. The study can help to develop products for the burned areas detection on a local scale since the fusion of one or more products in a grid can produce data closer to the real burned area. However, here we tested for a local at the Amazon and the product performance may vary according to other ecosystems. It is desirable to have an independent burned area map, produced with detail and human editing for a more comprehensive quantification of the operational product strengths and limitations.

Funding: This research was funded by the Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant number 2020/08916-8, 2020/15230-5, Fundação de Amparo à Pesquisa do Estado do Amazonas (FAPEAM) grant number 0102016301000289/2021-33 and the National Council for Scientific and Technological Development (CNPq), process number 314473/2020-3.

6. References

- Alencar, A. A. C., Arruda, V. L. S., Vieira, W., Conciani, D. E., Costa, D. P., Crusco, N., Galano, D. et al. (2022) “Long-Term Landsat-Based Monthly Burned Area Dataset for the Brazilian Biomes Using Deep Learning”, In *Remote Sensing*, vol 14, page 29.
- Alonso-Canas, I., and Chuvieco, E. (2015) “Global Burned Area Mapping from ENVISAT-MERIS and MODIS Active Fire Data”, *Remote Sensing of Environment*, vol 163 (June), pages 140–52.
- Alvares, C. A., Stape, J. L., Sentelhas, P. C., Gonçalves, J. L. M., and Sparovek, G. (2013) “Köppen’s Climate Classification Map for Brazil”, *Meteorologische Zeitschrift*, vol 22 (6), pages 711–28.
- Anderson, L. O., Aragão, L. E. O. C., Gloor, M., Arai, E., Adami, M., Saatchi, S. S., Malhi, Y. et al. (2015) “Disentangling the Contribution of Multiple Land Covers to Fire-Mediated Carbon Emissions in Amazonia during the 2010 Drought”. *Global Biogeochemical Cycles*, vol 29 (10), pages 1739–53.
- Andrade, D. F. C., Ruschel, A. R., Schwartz, G., Carvalho, J. O. P., Humphries, S. and Gama, J. R. V. (2020) “Forest Resilience to Fire in Eastern Amazon Depends on the Intensity of Pre-Fire Disturbance”. *Forest Ecology and Management*, vol 472(September), page 118258.
- Aragão, L. E.O.C., Anderson, L. O., Fonseca, M. G., Rosan, T. M., Vedovato, L. B., Wagner, F. H., Silva, C.V.J. et al. (2018) “21st Century Drought-Related Fires Counteract the Decline of Amazon Deforestation Carbon Emissions” *Nature Communications*, vol 9 (1), pages 1–12.
- Aragão, L. E. O. C., Silva-Junior, C.H. L., and Anderson, L. O. (2020). “O desafio do Brasil para conter o desmatamento e as queimadas na amazônia durante a pandemia por COVID-19 Em 2020: Implicações Ambientais, Sociais e Sua Governança.”. São José dos Campos..
- Arruda, V. L.S., Piontekowski, V. J., Alencar, A., Pereira, Reginaldo S., and Matricardi, E. A.T. (2021) “An Alternative Approach for Mapping Burn Scars Using Landsat Imagery, Google Earth Engine, and Deep Learning in the Brazilian Savanna” *Remote Sensing Applications: Society and Environment*, vol 22, page 100472.
- Assis, L. F. F. G., Ferreira, K. R., Vinhas, L., Maurano, L., Almeida, C., Carvalho, A., Rodrigues, J., Maciel, A. and Camargo, C. (2019) “TerraBrasilis: A Spatial Data Analytics Infrastructure for Large-Scale Thematic Mapping” *ISPRS International Journal of Geo-Information*, vol 8 (11), pages 513.
- Barlow, J., Berenguer, E., Carmenta, R. and França, F. (2020) “Clarifying Amazonia’s Burning Crisis” *Global Change Biology*, vol 26 (2), page 319–21.
- Barni, P. E., Pereira, V. B., Manzi, A. O. and Barbosa, R. I. (2015) “Deforestation and Forest Fires in Roraima and Their Relationship with Phytoclimatic Regions in the Northern Brazilian Amazo.” *Environmental Management*, vol 55 (5), pages 1124–38.
- Berlinck, C.N, and Batista, E. K. L. (2020) “Good Fire, Bad Fire: It Depends on Who Burns” *Flora Morphol. Distrib. Funct. Ecol. Plants*, vol 268, page 151610.
- Bivand, R. and Rundel, C. (2018) “Rgeos: Interface to Geometry Engine—Open Source (‘GEOS’)”. <https://rdr.io/cran/rgeos/>, May.

- Boschetti, L., Sparks, A., Roy, D. P., Giglio, L. and San-Miguel-Ayanz, J. (2020) “GWIS National and Sub-National Fire Activity Data from the NASA MODIS Collection 6 Burned Area Product in Support of Policy Making, Carbon Inventories and Natural Resource Management”. NASA Applied Sciences Grant USA. <https://gwis.jrc.ec.europa.eu/apps/country.profile/downloads>, May.
- Bush, M. B., Silman, M. R., McMichael, C. and Saatchi, S. (2008) “Fire, Climate Change and Biodiversity in Amazonia: A Late-Holocene Perspective” *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol 363 (1498), pages 1795–1802.
- Campanharo, W. A., Morello, T., Christofoletti, M. A. M. and Anderson, L.O. (2022) “Hospitalization Due to Fire-Induced Pollution in the Brazilian Legal Amazon from 2005 to 2018.” *Remote Sensing*, vol 14 (1).
- Carvalho, N. S., Anderson, L. O., Nunes, C. A., Pessoa, A. C. M., Silva Junior, C. H. L., Reis, J. B. C., Shimabukuro, Y. E., Berenguer, E., Barlow, J. and Aragao, L. E. O. C. (2021) “Spatio-Temporal Variation in Dry Season Determines the Amazonian Fire Calendar”. *Environmental Research Letters*, vol 16 (12).
- Chuvieco, E., Lizundia-Loiola, J., Pettinari, M. L., Ramo, R., Padilla, M., Tansey, Kevin, Mouillot, Florent et al (2018) “Generation and Analysis of a New Global Burned Area Product Based on MODIS 250 m Reflectance Bands and Thermal Anomalies”. *Earth System Science Data*, vol 10 (4), pages 2015–31.
- Dinamica EGO Team (2020) “*Calc Reciprocal Similarity Map*.” https://csr.ufmg.br/dinamica/dokuwiki/doku.php?id=calc_reciprocal_similarity_map, May.
- Fearnside, P. M. (2008) “Amazon Forest Maintenance as a Source of Environmental Services” *Anais Da Academia Brasileira de Ciências*, vol 80 (1), pages 101–14.
- Giglio, L., Boschetti, L., Roy, D. P., Humber, M. L. and Justice, C. O. (2018) “The Collection 6 MODIS Burned Area Mapping Algorithm and Product” *Remote Sensing of Environment*, vol 217 (July), pages 72–85.
- Hijmans, R. J. (2017) “*Raster: Geographic Data Analysis and Modeling*” <https://rdr.io/cran/raster/>, May.
- Humber, M. L., Boschetti, L., Giglio, L. and Justice, C.O. (2019) “Spatial and Temporal Intercomparison of Four Global Burned Area Products” *International Journal of Digital Earth*, vol 12 (4), pages 460–84.
- Justice, C. O., Giglio, L., Korontzi, S., Owens, J., Morisette, J. T., Roy, D., Descloitres, J., Alleaume, S., Petitcolin, F. and Kaufman, Y. (2002) “The MODIS Fire Products.” *Remote Sensing of Environment*, vol 83 (1–2), pages 244–62.
- Langford, Z., Kumar, J. and Hoffman, F. (2019) “Wildfire Mapping in Interior Alaska Using Deep Neural Networks on Imbalanced Datasets”. *IEEE International Conference on Data Mining Workshops, ICDMW*, vol 2, pages 770–78.
- Leite-Filho, A. T., Soares-Filho, B. S., Davis, J. L. and Rodrigues, H. O. (2020). “*Modeling Environmental Dynamics with Dinamica EGO*.” https://www.csr.ufmg.br/dinamica/dokuwiki/doku.php?id=guidebook_start, May.
- Leite-Filho, A. T., Soares-Filho, B. S., Davis, J. L., Abrahão, G. M. and Börner, J. (2021) “Deforestation Reduces Rainfall and Agricultural Revenues in the Brazilian Amazon”. *Nature Communications*, vol 12 (1), pages 1–7.

- Long, T., Zhang, Z., He, G., Jiao, W., Tang, C., Wu, B., Zhang, X., Wang, G and Yin, R. (2019) “30m Resolution Global Annual Burned Area Mapping Based on Landsat Images and Google Earth Engine.” *Remote Sensing*, vol 11 (5), pages 1–30.
- Maraseni, T. N., Reardon-Smith, K., Griffiths, G. and Apan, A. (2016) “Savanna Burning Methodology for Fire Management and Emissions Reduction: A Critical Review of Influencing Factors” *Carbon Balance Manag*, vol 1.
- Mataveli, Guilherme A.V., Michel E.D. Chaves, Nathaniel A. Brunsell, and Luiz E.O.C. Aragão. 2021. “The Emergence of a New Deforestation Hotspot in Amazonia.” *Perspectives in Ecology and Conservation* 19 (1): 33–36. <https://doi.org/10.1016/j.pecon.2021.01.002>.
- Mataveli, G.A.V., Gabriel de Oliveira, Hugo T. Seixas, Gabriel Pereira, Scott C. Stark, Luciana V. Gatti, Luana S. Basso, et al. 2021. “Relationship between Biomass Burning Emissions and Deforestation in Amazonia over the Last Two Decades.” *Forests* 12 (9): 1–19. <https://doi.org/10.3390/f12091217>.
- Mataveli, G. A. V., Silva, M. E. S., Pereira, G., Cardozo, F. S., Kawakubo, F. S., Bertani, G., Costa, J. C., Ramos, R. C., and da Silva, V. V. (2017) “Analysis of Fire Dynamics in the Brazilian Savannas”. *Natural Hazards and Earth System Sciences Discussions*, no. March, pages 1–27.
- Meddens, A. J. H., Kolden, C. A., Lutz, J. A., Abatzoglou, J. T. and Hudak, A. T. (2018) “Spatiotemporal Patterns of Unburned Areas within Fire Perimeters in the Northwestern United States from 1984 to 2014”. *Ecosphere* 9, page e02029.
- Mendonça, M. J. C., Diaz, M. C. V., Nepstad, D., Motta, R. S., Alencar, A., Gomes, J. C. and Ortiz, R. A. (2004) “The Economic Cost of the Use of Fire in the Amazon”. *Ecological Economics*, vol 49 (1), pages 89–105.
- Morton, D. C., DeFries, R. S., Nagol, J., Souza, C. M., Kasischke, E. S., Hurtt, G. C., and Dubayah, R. (2011) “Mapping Canopy Damage from Understory Fires in Amazon Forests Using Annual Time Series of Landsat and MODIS Data”. *Remote Sensing of Environment*, vol 115 (7), pages 1706–20.
- Mouillot, F., Schultz, M. G., Yue, C., Cadule, P., Tansey, K., Ciais, P. and Chuvieco, E. (2014) “Ten Years of Global Burned Area Products from Spaceborne Remote Sensing—A Review: Analysis of User Needs and Recommendations for Future Developments”. *International Journal of Applied Earth Observation and Geoinformation*, vol 26 (February), pages 64–79.
- Padilla, M., Stehman, S. V., Ramo, R., Corti, D., Hantson, S., Oliva, P., Itziar Alonso-Canas, et al. 2015. “Comparing the Accuracies of Remote Sensing Global Burned Area Products Using Stratified Random Sampling and Estimation.” *Remote Sensing of Environment* 160 (April): 114–21. <https://doi.org/10.1016/j.rse.2015.01.005>.
- Penha, T. V., Körting, T. S., Fonseca, L. M. G., Silva, C. H. L., Pletsch, M. A. J. S., Anderson, L. O. and Morelli, F. (2020) “Burned Area Detection in the Brazilian Amazon Using Spectral Indices and GEOBIA”. *Revista Brasileira de Cartografia*, vol 72 (2), pages 253–69.
- Pessôa, A. C; M., Anderson, L. O., Carvalho, N. S., Campanharo, W. A., Silva Junior, C. H. L., Rosan, T. M., Reis, J. B. C. et al (2020) “Intercomparison of Burned Area Products and Its Implication for Carbon Emission Estimations in the Amazon”.

Remote Sensing, vol 12 (23), page 3864.

- Prentice, I. C., Kelley, D. I., Foster, P. N., Friedlingstein, P., Harrison, S. P. and Bartlein, P. J. (2011) “Modeling Fire and the Terrestrial Carbon Balance”. *Global Biogeochemical Cycles*, vol 25 (3).
- Qgis (2019) “*QGIS Geographic Information System*”. Open Source Geospatial Foundation Project. https://qgis.org/pt_BR/site/, May.
- R Core Team (2021) “*R: A Language and Environment for Statistical Computing 2020*”. <https://www.r-project.org/>.
- Rodrigues, J. A., Libonati, R., Pereira, A. A., Nogueira, J. M. P., Santos, F. L. M., Peres, L. F., Rosa, A. S., et al. (2019) “How Well Do Global Burned Area Products Represent Fire Patterns in the Brazilian Savannas Biome? An Accuracy Assessment of the MCD64 Collections”. *International Journal of Applied Earth Observation and Geoinformation*, vol 78, pages 318–31.
- Roy, D.P. and Boschetti, L. (2009) “Southern Africa Validation of the MODIS, L3JRC, and GlobCarbon Burned-Area Products”. *IEEE Transactions on Geoscience and Remote Sensing*, vol 47 (4), pages 1032–44.
- Safi, Y. and Bouroumi, A. (2013) “Prediction of Forest Fires Using Artificial Neural Networks”. *Applied Mathematical Sciences*, vol 7 (5–8), pages 271–86.
- Shakesby, R. A. and Doerr, S. H. (2006) “Wildfire as a Hydrological and Geomorphological Agent”. *Earth-Sci. Rev.*, vol 74, pages 269–307.
- Shimabukuro, Y. E., Miettinen, J., Beuchle, R., Grecchi, R. C., Simonetti, D. and Achard, F. (2015) “Estimating Burned Area in Mato Grosso, Brazil, Using an Object-Based Classification Method on a Systematic Sample of Medium Resolution Satellite Images”. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol 8 (9), pages 4502–8.
- Silva Junior, C. H. L., Anderson, L. O., Silva, A. L., Almeida, C. T., Dalagnol, R., Pletsch, M. A. J. S., Penha, T. V., Paloschi, R. A. and Aragão, L. E.O.C. (2019) “Fire Responses to the 2010 and 2015/2016 Amazonian Droughts”. *Frontiers in Earth Science*, vol 7, pages 1–16.
- Smirnov, N.V (1939) “Estimate of Deviation between Empirical Distribution Functions in Two Independent Samples”. *Bull. Math. Univ. Moscou*, vol 2, pages 3–14.

Generating POI type embeddings based on variable Word2Vec sentences

Salatiel Dantas Silva¹, Claudio E. C. Campelo¹, Maxwell Guimarães de Oliveira¹

¹Systems and Computing Department (DSC)
Federal University of Campina Grande (UFCG)
Av. Aprígio Veloso 882, Bloco CN, Bairro Universitário – 58.429-140
Campina Grande – PB – Brazil

{salatiel@copin, {campelo, maxwell}@computacao}.ufcg.edu.br

Abstract. *Point of Interest (POI) types are one of the most researched aspects on urban data. Several works have successfully modeled POI types considering POI co-occurrences in different spatial regions along with statistical models based on the Word2Vec technique from Natural Language Processing. Although these works have presented good results, they do not consider the spatial distance among related POIs as a feature to represent POI types. In this context, we present an approach based on Word2Vec sentences that includes such a distance to generate POI type embeddings, providing an improved POI type representation. Experiments based on similarity assessments between POI types revealed that our representation provides values close to human judgment.*

1. Introduction

The importance of Points of Interest (POIs) for human activities has given rise to several kinds of researchs concerning reasoning and representing their types (such as *Restaurants* and *Parks*). Developing new methods capable of capture the semantics and similarity of POI types enables the creation of computational mechanisms that may assist in tasks such as urban planning and POI recommendation. Recent researches [Harispe et al. 2015, Yan et al. 2017, Yao et al. 2017] proposed capture the semantics and similarity of POI types considering spatial information that implicitly relates the types, such as POI co-occurrences in different spatial regions. For example, we can say that the POI type *Park* is somehow similar to the POI type *Playground* because they appear in similar contexts (they usually have similar vicinity). Hence, it is possible to calculate the similarity between POI types based on their contexts.

In this regard, Place2Vec [Yan et al. 2017] is one of the pioneers representing POI types using co-occurrence of POIs in a region along with the Word2Vec model [Mikolov et al. 2013]. In Word2Vec, words in a text are represented as real values' vectors (named word embeddings, vector embeddings or embeddings) based on the co-occurrence of the words [Bengio et al. 2003, Mikolov et al. 2013]. Concerning POIs, Place2Vec [Yan et al. 2017] applied this technique to extract and measure the relation of POI types in a region according to the relation between its vectors (called POI type embeddings). Thus, POI types with similar semantics will be closer in a vector space. These vectors can be used as input on clustering or recommendation algorithms, allowing tasks such as POI recommendation and POI type automated classification, that may contribute to improve POI-based urban planning, bussiness analysis, and so on.

Even though Place2Vec has presented good results, there are some limitations to be considered. Such a method does not consider the distance among POIs as a feature to capture their similarity. The distance was only considered to check which POIs are in a given vicinity. However, according to the Tobler’s first law of geography: “Everything is related to everything else, but near things are more related than distant things” [Tobler 1970]. Thus, we agree that the closer are the POIs the stronger is the relationship among them. Another aspect worth mentioning is that the POI types have different levels arranged in a hierarchy¹. Nonetheless, in Place2Vec, all POI types are related, with no distinction between different levels. Hence, types of a certain level can be directly related to types of other levels. This approach does not allow models to capture the semantics of types for each level separately, preventing comparisons across levels more accurately.

On this basis, this paper presents the Sentence Information Theoretic Distance Lagged (S-ITDL), an alternative to generate POI type embeddings that considers the distance between POIs in a vicinity to capture their similarity. Our approach uses the distance to adjust the training weights of each POI type using Word2Vec’s architecture. Furthermore, we generated POI type embeddings for the various levels of the hierarchy to understand how they impact the POI type semantics. The main contributions of this article are:

- A new method to represent POI types integrating the distance between POIs as a weight adjustment factor in the training of models;
- An evaluation of the POI type embeddings based on the hierarchy, analyzing each taxonomy levels separately;
- A Github repository containing the developed code, train and test set, and the trained models².

We organized the subsequent sections in this paper as follows. Section 2 reports a range of works developed in research that focus on the computational representation of POI types. After that, Section 3 presents our approach, exposing how we applied POI distance to update weights in the training of POI types. Following this, Section 4 provides an overview of the data used in the models’ training. Afterward, Section 5 addresses the evaluation setup and the experimental results, as well as comparisons between our models’ results and the baseline. Lastly, Section 6 summarizes the results found, indicating some future research possibilities.

2. Related Work

This section discusses how related works represent POI types using computational mechanisms.

2.1. Latent Representation

The Word2Vec, proposed by [Mikolov et al. 2013], is a technique developed for Natural Language Processing capable of generating predictive models from raw text. Word2Vec contains two architectures, as illustrated in Figure 1. In the Continuous Bag of Words (CBOW), given the preceding word sequence $w(k-1)$, $w(k-2)$, and the succeeding

¹ POI level hierarchy: https://www.yelp.com/developers/documentation/v3/category_list

² Source code: <https://github.com/SalatielDantas/S-ITDL.git>

word sequence $w(k+1)$, $w(k+2)$, the model can calculate the probability of a word w_k appear. In the Skip-Gram architecture, the model must predict the context words (preceding and succeeding) from the center word (w_k). Word2Vec model uses a window with variable size to define the “context words”. For example, if the window size is five, the context is the five words that precede a target word and the five words that succeed it.

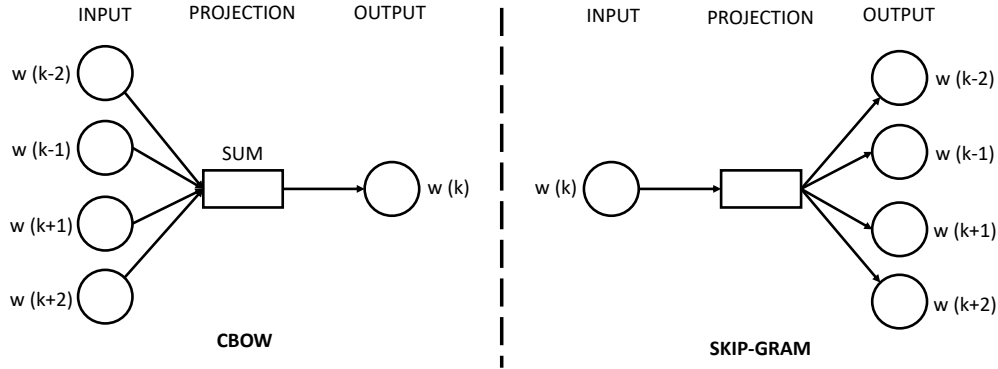


Figure 1. Word2Vec’s Neural Network Architectures [Mikolov et al. 2013]

In its architecture, Word2Vec has a hidden layer (projection) that learns each word’s embedding, yielding a d -dimensional embedding space. Thus, one can use the hidden layer weights (vector) for each word and apply vector space operations, such as cosine similarity. In this way, the similarity between different words can be calculated, since words with similar co-occurring contexts will have similar embeddings.

In the POI scenario, the idea is similar. The model predicts a center POI type with a set of context POI types (from the vicinity) or vice-versa. Thus, we can define a language model for POI types and use it’s vectors to calculate POI type similarity.

2.2. POI Type Embeddings

As an alternative to a computational representation of POI types, in Place2Vec, [Yan et al. 2017] proposed the Information Theoretic, Distance Lagged (ITDL) algorithm aiming to capture the POI types’ semantics and, as a consequence, understand their similarity and relationship. This method considers that the continuous vicinity of a POI could be divided into multiple discrete bins (which are stripes of space around a POI). Figure 2 illustrates this configuration. Each bin represents a region used to learn a latent representation and encodes the semantic distribution between a center POI and its neighbors (also called context POIs). Two main pieces of information are considered to model the similarity and relatedness between the center POI Type and context POI Type: human activity (represented by the check-ins count of POIs in a bin) and the uniqueness of a place (represented by the probability of a POI appearing in a bin). They are combined to generate a number that is used to augment the relation $\langle \text{center POI type, context POI type} \rangle$ in the train set.

The algorithm uses binary relations between the center POI and the context POIs in each bin generating a list of tuples in the format $\langle \text{center POI type, context POI type} \rangle$. This set of tuples can be interpreted as a two-word statements, which is suitable for input into Word2Vec. During training, Word2Vec estimates the probability



Figure 2. An example of vicinity view from the discrete bins.

that a context POI Type occur given the central POI type. Thus, each POI type is represented by the co-occurrences with other POI types in a bin.

Similarly, Liu et al. [Liu et al. 2019] applied Word2Vec to extract and represent a place’s “niche” patterns. The developed models generate two main results: representations for the place’s niche by type in a latent vectorial space, by which nearby vectors represent similar niches; and the occurrence probability of each POI type in the vicinity of a central POI.

The POI type embeddings generated by Yan et al. [Yan et al. 2017] were applied in later works by different approaches. Liu et al. [Liu et al. 2020] proposed a framework to envision and explore POIs, using dimensionality reduction techniques, such as t-SNE [Van der Maaten and Hinton 2008]. In this framework, the POI vectors are mapped to a two-dimensional space. Then, the configuration of POIs’ regions is rendered by a thematic map that enables comparisons between different locations.

Zhai et al. [Zhai et al. 2019] also proposed a framework based on ITDL to extract and identify the functions of urban regions. In this approach, the authors add the vectors of POIs in a vicinity to represent as a single entity. Next, they applied the K-Means method to group similar entities. Furthermore, they enriched each entity by adopting vehicle and people traffic information to assist in tasks such as urban planning.

Hu et al. [Hu et al. 2020] also applied the idea of POIs’ co-occurrence in a vicinity, using vector embedding multiprototype to encode the POI and topics that characterize it at once. After encoding, the vectors were clustered using the HDBSCAN algorithm [McInnes et al. 2017] to generate POIs’ regions of similar types and characteristics.

3. Sentence-ITDL

As mentioned in Subsection 2.2, The ITDL algorithm creates a set composed by tuples in the format $\langle \text{center POI type, context POI type} \rangle$ that can be used as a two-word phrase in Word2Vec’s training. Figure 3 illustrates an example of training a set of tuples in Word2Vec’s Skip-gram architecture. The first word in each tuple represents the center POI type. The second word, represents the context POI type. On the right side of Figure 3, is the tuples processing. In each instant $t_1, t_2, \dots, t_n$, the weights (w') of each relation $\langle \text{center POI type, context POI type} \rangle$ are adjusted to estimate the probability of occurrence of the context POI type ($w(k+1)$), given the center POI type ($w(k)$).

Our approach also follows a similar idea, but including the distance to adjust the

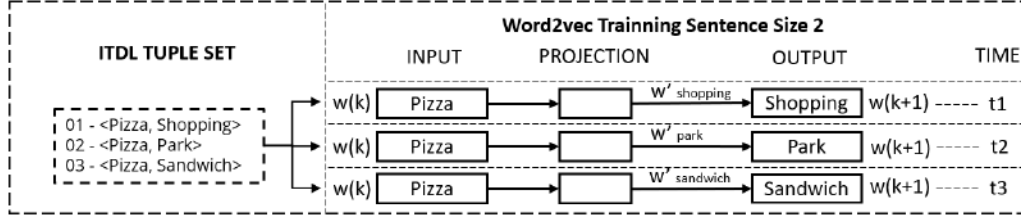


Figure 3. ITDL training example using Word2Vec's Skip-Gram architecture

weights of each sentence in Word2Vec. The penalization of the POI types relationship is applied using Word2Vec's structure. As seen in Figure 1, it is possible to train the center word ($w(k)$) considering the words strictly next to it ($w(k+1)$, $w(k-1)$) and the ones further away ($w(k+2)$, $w(k-2)$). In Skip-Gram, an intrinsic penalty is applied on the weights as the words move away from the center word [Mikolov et al. 2013]. The more distant the context word, the more penalized it will be in the co-occurrence relationship. Based on this structure, we formulated the Sentence-ITDL (S-ITDL) algorithm that converts tuples into sentences with variable sizes, using the distance between POIs to define the types' position in the generated sentences. Then, when we train the model, the position of each POI type in the sentence, now defined by the distance to the center POI, will affect the estimation of its types relation weight (w').

3.1. Tuples to Sentence With Variable Size

To make the conversion of tuples into sentences, we have made a change in the ITDL algorithm. The tuples generated are saved in the database along with the distance between the center POI and the context POI, resulting in a triple $\langle \text{center POI type, context POI type, distance} \rangle$. Once the build of the dataset is ready, it goes through another process: converting it into sentences with a pre-defined size. Then, we use the new dataset of sentences in the training process. Figure 4 illustrates the process followed by our approach.

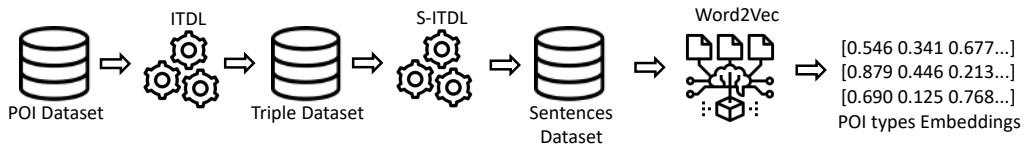


Figure 4. Process to generate embeddings with the ITDL and S-ITDL algorithms

The first step consists in executing the ITDL algorithm along with some POI dataset. The second step is the execution of S-ITDL, which transforms the triples into sentences. At the end of the process, the Word2Vec is trained using the sentences. The S-ITDL algorithm, presented in Algorithm 1, receives as input a set of triples (T), consisting of a center POI type ($tp1$), a context POI type ($tp2$), the distance (d) among these POIs, the distance from the center POI to the discrete bin (b) and the size of the sentences (s). In the algorithm, the triples set (T) is given to the `group_by_type` function (line 2) that groups the POIs types concerning the center POI types, ordered from the spatially closest to the farthest. To insert the POI types in a sentence, we fixed the first position to put the center POI type while the other positions of a sentence contain the context POI types. POIs are distributed in a continuous geographic space, and it is not possible to insert them

directly into sentences that have a fixed size. For this we define that each sentence index will be mapped to space intervals to cover the POIs. For that, we define a compartment size (c) that map each index to a range of spaces to the sentences index.

Algorithm 1: S-ITDL ALGORITHM

Input: $T = (tp1, tp2, d), b, s$
Output: List of sentences

```

1 sentences  $\leftarrow$  empty_list()
2  $G \leftarrow$  group_by_type( $T$ )
3  $c \leftarrow b/s$ 
4  $s \leftarrow s + 1$ 
5 foreach  $g \in G$  do
6    $i \leftarrow 0$ 
7   while  $i < b$  do
8      $K \leftarrow$  types_in_range( $g, i, c$ )
9     foreach  $k \in K$  do
10      subsentence  $\leftarrow$  list_with_Sentinel()
11      subsentence[0]  $\leftarrow$  k['center_POI_type']
12      subsentence[i/c]  $\leftarrow$  k['context_POI_type']
13      sentences.append(subsentence)
14    $i \leftarrow i + 1$ 
15 return sentences

```

In order to fill the sentences, we traverse distance intervals starting from the location of center POI and ending at b meters (the last bin distance), with the addition of c meters between each interval. The `types_in_range` function captures the POI types in the interval of distances greater than i and smaller than $i + c$. In other words, the function searches for POIs that fit in a range mapped to a specific sentence index. When identifying all POIs in the interval, the algorithm loops through one by one, creating sub-sentences. This sub-sentence is filled with a ‘Sentinel’ word, indicating that certain position do not contain POI types. Then the center POI type is placed at the beginning of the sentence (index 0), and the context POI type at index (i/c). Finally, the new subsentence is added to the list of sentences. It is important to highlight that each subsentence consists only of a center POI type and a context POI type, with the aim of maintain the binary relation generated by the ITDL. The insertion of a third POI type, for instance, would modify such a relationship in a subsentence

As an example, the left part of Figure 5 illustrates a situation in which the distance of the discrete bin (b) used is 100m. The sentence size (s) is 2, indicating that the sentence will have two context words and a center word (index 0). The compartment size(c) is 50m (b/s). The center POI type will fill the first sentence index. The context POI types in $[0, 50]$ meters range will fill the second index. The context POI types in $[51, 100]$ meters range will fill the last index. Firstly, the triples 01 and 02 are converted for being in the interval $[0, 50]$. For these two sentences, the last index is filled with ‘Sentinel’ word. We do not fill with another POI type to in the empty indexes because we aim to keep the binary relation between the center POI and context POI individually. Then, we updated the range to $[51, 100]$ meters and converted the triple 03.

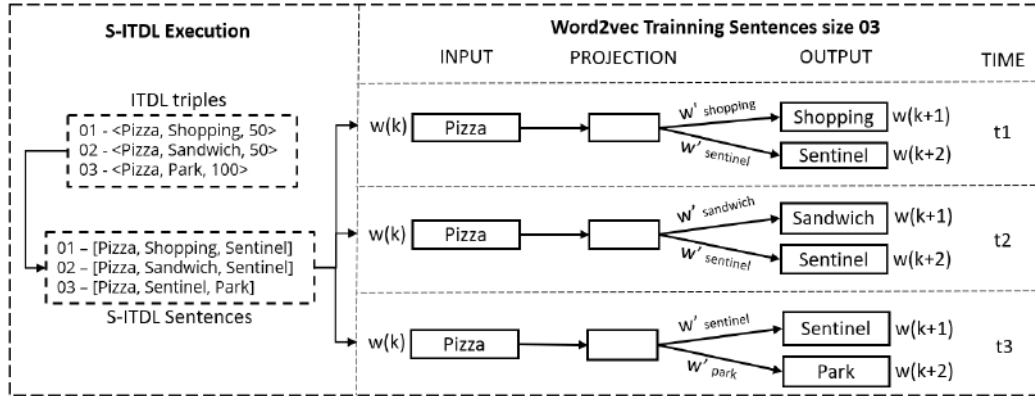


Figure 5. Example of S-ITDL sentences training in the Word2Vec's Skip-Gram architecture

The right side of Figure 5 shows how training occurs in Word2Vec. The POIs closer to the center are less penalized than those further away. In the example, the Shopping and Sandwich types are trained at position $w(k+1)$, while the Park POI type, which is more distant, is trained at position $w(k+2)$. Therefore, the weight of relationship $\langle \text{Pizza}, \text{Park} \rangle$ is more penalized than the other weights relationships.

During the training step, we set the Word2Vec to ignore the ‘Sentinel’ word, as its function is only to serve as a word that fills an empty space in the sentence. With this new approach, we can incorporate distance information among POIs to generate POI type embeddings capable of capturing the distance aspects among POIs that keeps the types.

4. Dataset

The POIs used in this paper come from the Yelp Challenge³ database (February 2021 version). We chose this database because it was used [Yan et al. 2017], our baseline. It comprises 160,585 POIs, spread across 836 cities in the USA and Canada. Our experiments used the Austin region because it covers all POIs present in the evaluation set.

Each POI is composed of 14 attributes. In our experiments, we used: business_id, a unique identifier; latitude and longitude, indicating the exact location of the POI; categories which indicates the POI types; and the check-in count that each POI presents, as this information is needed to run the ITDL and S-ITDL algorithms. Figure 6 illustrates the POI structure in Yelp dataset. Furthermore, Yelp has a taxonomy for the types of places, grouping them into 21 root types (level 01), 857 types on level 02, 404 types on level 03 and 17 types on level 04.

5. Experiments and Results

This section describes the experimental evaluation we conducted.

5.1. Evaluation Schema

The set of experiments we conducted was the same performed by the baseline research [Yan et al. 2017], since the aim is to provide a direct comparison to it. The first

³<https://www.yelp.com/dataset>

	business_id	latitude	longitude	categories	checkin_count
0	bBDDEgkFA1Otx9Lfe7BZUQ	36.208102	-86.768170	Ice Cream & Frozen Yogurt, Fast Food, Burgers,...	21
1	4iRzR7OaS-QaSxuvYxEGKA	36.248570	-86.719986	Hot Dogs, Restaurants	39
2	tMkwHmWFUEXrC9ZduonpTg	36.159886	-86.773197	Restaurants, Japanese, Seafood	96
3	Ik9IwjZXqUMqQOhM774DtQ	36.148371	-86.798895	Coffee & Tea, Restaurants, Wine Bars, Bars, Ni...	204
4	B2qjCl_BNNnXvWWfpFiLA	36.103614	-86.816268	Shopping, Baby Gear & Furniture, Home Decor, F...	13

Figure 6. Yelp dataset example

method, named Binary HIT Evaluation (BHE), consists in identifying the less similar POI Type, taking into account three different POI types disposed in a triple. In this task were presented 80 triples, and for each one, the participants chose the least similar POI Type between them.

The second method, named Ranking-based HIT Evaluation (RHE), aimed to indicate the relationship of similarity between two different POI Types. The higher the RHE value, the greater the similarity relationship. Based on these values, was calculated the Spearman’s correlation coefficient [Ramsey 1989] for both people and models’ results. This coefficient indicates not only the strength but also the direction of the association that exists between two variables [Ramsey 1989]. This experiment used 70 pairs of POI Types. To complement this assessment, we used the Mean Square Error (MSE) calculation between the values given by the models and the values given by people to measure which ones best capture the intensity of the relationship between the POI Types identified by the participants.

On the set of tests, comparisons between types of different levels was observed. For instance, the Pubs type (level 03) was one of the test examples that was compared to the Cupcake type (level 02). It is important to highlight that we seek to evaluate the POI type embeddings for different levels of hierarchy to understand how each method captured the semantics at these levels. Therefore, some operations on the test set were applied. For each line of the set, we defined the most general type, then, the types compared to the same level were generalized. This process resulted in the following datasets separated by levels:

- BHE Level 01 (30 test triples);
- BHE Level 02 (42 test triples);
- RHE Level 01 (16 test tuple);
- RHE Level 02 (43 test tuple);
- RHE Level 03 (08 test tuple).

The new test datasets did not have level types greater than 03 since there is no type greater than this level in the original set. As a consequence, for BHE our assessments will be at levels 01 and 02; for RHE, our tests will be at levels 01, 02, and 03.

5.2. Parameters Configuration

To run the experiments, we must define parameters related to S-ITDL, Word2Vec, and ITDL. For S-ITDL, it is essential to define the size of the sentences to be created (s) and the bin distance (b) to the bin used. Regarding the ITDL parameters, it is necessary to determine the quantity of bins and the importance (w) of the uniqueness (given by the

probability of a POI appearing in a bin) and popularity (given by check-ins). The value w varies in $[0, 1]$. When $w = 0$, we considered only the check-ins of the POIs. On the other hand, when $w = 1$, we considered only the probability of POI occurrence in a bin. If $w = 0.5$, the algorithm considers the two elements balanced. In our experiments we settled the importance to $w = 0.5$. For the quantity of discrete bins and dimension of the vectors, we used the same configuration defined by our baseline: 26 bins and 70 for vector dimension. For the Word2Vec window size, we used $window\_size = 05$, as it is a value that presents good results in the Skip-gram architecture⁴. For the sentence size, the value was the same as in the Word2Vec window, since the sentences are trained with this window in Word2Vec. In addition, we performed a test with a slightly smaller sentence size (04) in order to evaluate the behavior in this situation. The S-ITDL sentence sizes were $s = \{4, 5\}$. We used the Euclidean metric to measure the distance among POIs. However, other distance metrics might be used.

5.3. Results

This section discusses the results of the experiments carried out, bringing with it a comparison between our approach and the baseline. Considering the models generated for each bin, we calculated the average (avg), standard deviation (std), lowest value (min) and highest value (max). We grouped the results according to the levels of the Yelp hierarchy.

The first set of assessments is related to the RHE task. For this evaluation, we calculated the cosine similarity between pairs of vectors compared them to people's responses. In this task, we used the Average Spearman's Rank Correlation Coefficient metric (Spearman's ρ) as measure. This measure indicates the existing relationship between ranks [Ramsey 1989]. Their values are in the interval of $[-1, 1]$. Values close to 1 indicate a strong positive correlation, while values close to -1 indicate a strong negative correlation. If there is no correlation, the value approaches 0. Table 1 illustrates the results obtained between the results of the people and the models, where it is possible to see that the S-ITDL (size 04) were superior at level 02, with a avg value equal to 0.49 of Spearman's ρ . For level 03 the S-ITDL (size 05) were superior with a avg value equal to 0.77. For level 01 the results of ITDL was better. One hypothesis that may contribute to this result is the reduced number of types at this level and the high probability of these types co-occurring in a bin, generating high similarity values. For this reason, the penalty used in the S-ITDL caused the results to be below the ITDL. Although S-ITDL has not surpassed ITDL in level 01, its results show that at the other levels, where there are the most significant amount of types (see Section 4), pointing out that S-ITDL captures more accurately the semantics of POI types than the ITDL.

To complement this task, we calculated the Mean Squared Error (MSE) between the values given by the people and the values given by the models. It is important to notice that the smaller, the better the result. The results of this evaluation are in Table 2. Our models outperform the ITDL at all levels. This indicates that the S-ITDL models have greater precision in indicating the value of the relationship between the POI types than the ITDL.

The second set of assessments is related to the BHE task. In this task, we calcu-

⁴<https://code.google.com/archive/p/word2vec/>

Table 1. Spearman’s ρ RHE

	Level 01	Level 02	Level 03
ITDL			
avg	0.60	0.33	0.75
std	0.06	0.06	0.14
min	0.47	0.24	0.43
max	0.72	0.51	0.98
S-ITDL (size 04)			
avg	0.48	0.49	0.75
std	0.06	0.07	0.15
min	0.36	0.35	0.33
max	0.63	0.68	0.98
S-ITDL (size 05)			
avg	0.47	0.48	0.77
std	0.06	0.06	0.13
min	0.37	0.34	0.45
max	0.59	0.64	0.98

Table 2. MSE for RHE

	Level 01	Level 02	Level 03
ITDL			
avg	0.50	0.36	0.23
std	0.02	0.03	0.04
min	0.45	0.28	0.18
max	0.53	0.39	0.36
S-ITDL (size 04)			
avg	0.34	0.28	0.22
std	0.01	0.01	0.03
min	0.32	0.24	0.17
max	0.39	0.30	0.32
S-ITDL (size 05)			
avg	0.34	0.28	0.22
std	0.02	0.01	0.03
min	0.33	0.24	0.16
max	0.40	0.30	0.32

lated the cosine similarity among the vector pairs of each tuple of the test set. We found the most similar tuple, and the place that did not belong to the tuple becomes the task’s answer. Table 3 illustrates the results obtained. For level 01, it is possible to see that the models have statistically the same results. We believe that this result is due to the reduced number of types in this level. For level 02, the avg of S-ITDL size 04 and 05 was over ITDL. It is possible to notice that in level 02, the maximum accuracy values are better in the S-ITDL models. It is important to mention that level 02 contains a greater variety of POI types (See Section 4), indicating that our technique better captures the similarity between this variety of types.

Table 3. Accuracy for BHE

	Level 01	Level 02
ITDL		
avg	0.97	0.64
std	0.01	0.05
min	0.97	0.55
max	1.00	0.74
S-ITDL (size 04)		
avg	0.97	0.68
std	0.00	0.04
min	0.97	0.57
max	0.97	0.79
S-ITDL (size 05)		
avg	0.97	0.68
std	0.01	0.05
min	0.97	0.57
max	1.00	0.76

6. Conclusion and Further Work

On capturing the semantics of POI types by learning vector embeddings and using them to reason about places has already shown good results in several works. In this paper, we demonstrated that incorporating the distance among POIs in this learning process enables, for some hierarchy levels, a more precise differentiation of each type of POI. We also demonstrated that using the POI distance in a spatial region allows to achieve better accuracy to distinguish the POI types. Furthermore, we show that using the distance also enables each POI type vector to present similarity values closer to the values of human judgment. Moreover, the application of our technique does not require much effort, using Word2Vec’s own natural structure, not requiring many extra computational resources.

We have also revealed that separately considering the levels of types enables capturing the POI types semantics without interfering among the levels. The results demonstrate that each level can present different configurations that influence the POI type semantics. Thus, using vector embeddings by hierarchy level allows one to perform more general or more specific operations. Furthermore, we demonstrate that S-ITDL performs better at levels with a greater variety of POI types, indicating that this method is suitable in these situations. For levels with a smaller amount of types, the S-ITDL still presents values of similarity closer to human judgment.

In the future, we intend to apply the distance value to directly penalize the co-occurrence relationship in the Word2Vec’s training set. In other words, instead of considering distance ranges mapped to indexes, we intend to use each distance to penalize the relationship between the POIs uniquely. In this way, different distances will have different influences on each POI.

Acknowledgments

This research had financial support granted by the Paraíba State Research Foundation (FAPESQ), grant #2021/3217.

References

- Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155.
- Harispe, S., Ranwez, S., Janaqi, S., and Montmain, J. (2015). Semantic similarity from natural language and ontology analysis. *Synthesis Lectures on Human Language Technologies*, 8(1):1–254.
- Hu, S., He, Z., Wu, L., Yin, L., Xu, Y., and Cui, H. (2020). A framework for extracting urban functional regions based on multiprototype word embeddings using points-of-interest data. *Computers, Environment and Urban Systems*, 80:101442.
- Liu, K., Yin, L., Lu, F., and Mou, N. (2020). Visualizing and exploring poi configurations of urban regions on poi-type semantic space. *Cities*, 99:102610.
- Liu, X., Andris, C., and Rahimi, S. (2019). Place niche and its regional variability: Measuring spatial context patterns for points of interest with representation learning. *Computers, Environment and Urban Systems*, 75:146–160.
- McInnes, L., Healy, J., and Astels, S. (2017). hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.
- Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- Ramsey, P. H. (1989). Critical values for spearman’s rank order correlation. *Journal of educational statistics*, 14(3):245–253.
- Tobler, W. R. (1970). A computer movie simulating urban growth in the detroit region. *Economic geography*, 46(sup1):234–240.
- Van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(11).
- Yan, B., Janowicz, K., Mai, G., and Gao, S. (2017). From itdl to place2vec: Reasoning about place type similarity and relatedness by learning embeddings from augmented spatial contexts. In *Proceedings of the 25th ACM SIGSPATIAL international conference on advances in geographic information systems*, pages 1–10.
- Yao, Y., Li, X., Liu, X., Liu, P., Liang, Z., Zhang, J., and Mai, K. (2017). Sensing spatial distribution of urban land use by integrating points-of-interest and google word2vec model. *International Journal of Geographical Information Science*, 31(4):825–848.
- Zhai, W., Bai, X., Shi, Y., Han, Y., Peng, Z.-R., and Gu, C. (2019). Beyond word2vec: An approach for urban functional region extraction and identification by combining place2vec and pois. *Computers, Environment and Urban Systems*, 74:1–12.

Feature extraction of spatial panel data with autoencoders for clustering the Brazilian agricultural diversity

Marcos A. S. da Silva¹, Leonardo N. Matos², Flávio E. de O. Santos²,
Márcia H. G. Dompieri³, Fábio R. de Moura⁴

¹Embrapa Coastal Tablelands, Av. Beira Mar, 3250, 49025-040, Aracaju, SE, Brazil

²Department of Computer Science – UFS, 49100-000, São Cristóvão, SE, Brazil

³Embrapa Territorial, Av. Soldado Passarinho, 303, 13070-115, Campinas, SP, Brazil

⁴Department of Economics – UFS, 49100-000, São Cristóvão, SE, Brazil

marcos.santos-silva@embrapa.br, leonardo@dcomp.ufs.br,
flavioemanuel1859@gmail.com, marcia.dompieri@embrapa.br,
fabiromoura@gmail.com

Abstract. *Brazilian agricultural production presents a high degree of spatial diversity, which challenges designing territorial public policies to promote sustainable development. This article proposes a new approach to cluster Brazilian municipalities according to their agricultural production. It combines a feature extraction mechanism using Deep Learning based on Autoencoders and clustering based on k-means and Self-Organizing Maps. We used the panel data from IBGE's annual estimates of Brazilian agricultural production between 1999 and 2018. Different structures of simple stacked undercomplete autoencoders were analyzed, varying the number of layers and neurons in each of them, including the latent layer. We evaluated the asymmetric exponential linear loss function to cope with the sparse data. The results show that in comparison with the ground truth adopted, the autoencoder model combined with the Self-Organizing Maps and the k-means algorithm presented a better result than the clustering of the raw data from the k-means, demonstrating the ability of simple stacked autoencoders to reduce the dimensionality and create a new space of features in their latent layer where the data can be analyzed and clustered. Although the general accuracy is low, the results are promising, considering that we can add new improvements to the Deep Clustering process.*

1. Introduction

At the landscape level, Brazilian agricultural output exhibits various degrees of variation [Sales and Rodrigues 2019]. Each region specializes in agricultural activities, from high-tech agribusiness to well-known smallholders [Teixeira and Ribeiro 2020], due to climate, water resources, geographic limitations, and past socioeconomic processes. Furthermore, this heterogeneity is visible at intra- and regional scales.

By grouping the areas according to their shared agricultural production, it is possible to identify the regional particularities that must be considered when creating territorial public policies. Silva et al. (2022) divided the Brazilian municipalities into eight

agricultural production diversity trend groups, each associated with a distinct degree of native vegetation alteration, using clustering analysis and feature engineering. Their research provided compelling evidence for the claim made by [Fatch et al. 2021] that low-diversified regions typically exhibit a low degree of sustainability.

[Silva et al. 2022] used the IBGE’s annual agricultural production estimates for the years 1999 to 2018 to determine a diversity index based on Shannon’s entropy for each category (animal herd, planted area with temporary crops, the production value for temporary and permanent crops, aquaculture, silviculture, vegetal extractivism, and animal), totaling eight variables for 20 years and 5570 municipalities. They then used a shallow learning technique to cluster the spatial panel data based on the Self-Organizing Map Artificial Neural Network in conjunction with k-means.

The raw spatial panel data used by [Silva et al. 2022] comprises 196 variables for 20 years. Adopting a feature engineering strategy decreased the number of variables from 196 to 8 and eliminated various data issues like a high number of zeros and values close to zero. Therefore, the research questions are: What would happen if we directly ran a clustering analysis on the unprocessed panel data? How could we replace the feature engineering employed by [Silva et al. 2022] with a Deep Learning feature extraction?

This paper used a Deep Learning strategy based on autoencoders to extract features from raw Brazilian agricultural spatial panel data to cluster the data. The clustering obtained by [Silva et al. 2022] serves as the ground truth. Over the data projected on the new feature space, we used two classical clustering algorithms, k-means and Self-Organizing Maps associated with the k-means.

The paper is organized as follows: section 2 presents a brief review on deep clustering with autoencoders; section 3 discuss about the dataset and the proposed approach to feature extraction and spatial panel data clustering; section 4 shows the results and discussion; and section 5 unveil the conclusions.

2. Related work

Deep learning is a consolidating field in the industry, responsible for a significant transformation of data analytics, primarily in image, video, and text processing. Still, there are many research challenges, such as clustering using a deep learning technique known as deep clustering [LeCun et al. 2015].

In a detailed clustering investigation, [Min et al. 2018] identified various Deep Learning model architectures, mainly based on autoencoders. Autoencoders are deep artificial neural networks that use an unsupervised learning method to extract features from a dataset by combining an encoder (a nonlinear mapping function) and a decoder (a dataset reconstructor from the encoder-generated representation). A latent layer that stores the data representation can be used to help cluster the data while gathering all the information needed to reconstruct the entire dataset. Nevertheless, this latent layer representation does not preserve the original data topology and neighborhood properties.

According to [Min et al. 2018], there are at least two methods for performing Deep Clustering with autoencoders. First, use the autoencoder to reduce the dimensionality of the data before applying a clustering algorithm to the encoded data. Second, while the autoencoder parameters are being updated in the deep learning process,

combine the reconstruction loss with a clustering loss to cluster data [Charte et al. 2018, Song et al. 2014, Guo et al. 2017]. The Deep Clustering strategy is determined by dataset properties such as sparsity, size, format (images, sequences, and tabular), and structural complexity, and this will necessitate different approaches such as those proposed by [Du et al. 2021] for text and image clustering using a deep multi-view clustering algorithm based on multiple auto-encoders, [Xu et al. 2020] for image clustering using variational autoencoders, or [Falissard et al. 2018] for longitudinal tabular data combining recurrent neural network and autoencoders.

We concluded from the literature review that there are few works on tabular panel data, as in [Falissard et al. 2018], and that a clustering analysis with autoencoders implies an empirical data-driven process. Then, the most appropriate strategy for investigating how autoencoders can map the original tabular panel data to a new latent feature space is incrementally adding complexity to the model. Exploring data clustering directly from encoded data [Falissard et al. 2018], evaluating the combination of objective and clustering loss functions [Song et al. 2014], and finally testing more complex Deep Clustering propositions [Du et al. 2021, Xu et al. 2020].

3. Data and methods

3.1. Spatial panel data

The dataset comprises 196 variables of IBGE’s annual estimates for all Brazilian municipalities [IBGE 2021]. These variables correspond to eight groups: herd population, animal production value, planted temporary crops, silviculture, aquaculture, vegetal extractivism, and temporary and permanent crop production value. A detailed data description can be found in [Silva et al. 2022].

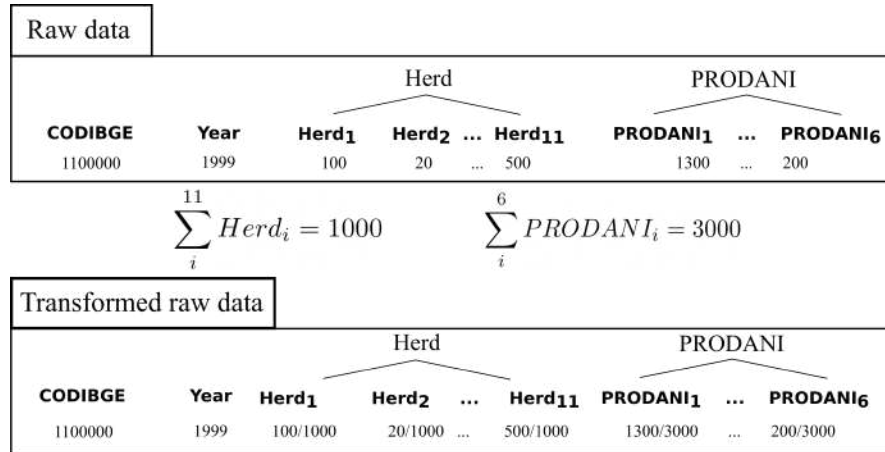


Figure 1. Example of raw data and its transformation. We converted each variable to its unit rate according to the category (e.g., herd or prodani) it belongs. Source: elaborated by the authors.

We transformed the raw data as follows: a) each variable is associated with only one category; b) for each observation (municipality-year), we calculate the sum for each category; c) each variable is updated by dividing its value by the sum of the category it belongs to. In the end, each variable will correspond to the unit rate of that product for

each observation (municipality-year) (Fig. 1). After that, we linearly normalized the data according to the min-max algorithm transforming all variables into the interval $[0, 1]$.

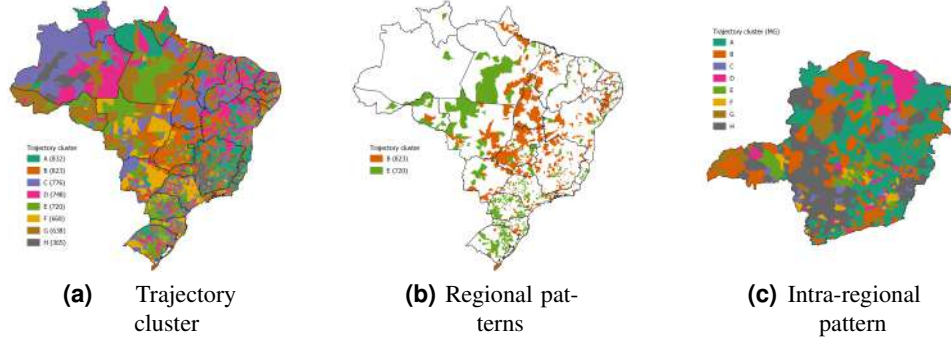


Figure 2. a) Clustered municipalities based on the trajectory of Shannon’s diversity indices onto a Self-Organizing Map by [Silva et al. 2022]. b) Mapping only two trajectory clusters to highlight regional patterns of clusters ‘B’ and ‘E’. c) The same clustering for the Minas Gerais state, showing intra-regional patterns. This trajectory cluster was considered the ground truth for this paper—source: elaborated by the authors.

The main characteristic of this dataset is the considerable presence of zeros. The mean and median percentages of zeros per variable in the entire dataset are 83.09% and 91.49%, respectively. They are structural zeros because most municipalities produce a limited amount of agricultural products, so no available data were set to zero. This unbalanced data challenges the learning process of artificial neural networks that will be induced to learn the zeros instead of the rest of the patterns. Initial tests showed that any autoencoder structure models used in this research converged for symmetric and some asymmetric (linear and quadratic) loss functions. Thus, this demanded investigating a suitable loss function for a very sparse dataset.

3.2. Proposed approach

The proposed approach comprises two steps. In the first step, we defined a set of autoencoder models (section 3.2.1), and performed a parameterization of an asymmetric loss function to cope with the data sparsity (section 3.2.2). The next step is responsible for clustering the encoded data onto a new and low dimensional feature space for each autoencoder model (section 3.2.3).

3.2.1. Autoencoder models

We have chosen six simple stacked undercomplete autoencoder models with the same optimizer (adam), hidden and output activation functions (relu and sigmoid), fully connected, with the same loss function, and varying the number of hidden layers and the size of the latent layer (Fig. 3). Table 1 shows the number of the hidden layers to the encoder component of each evaluated autoencoder, including the number of neurons on the latent hidden layer.

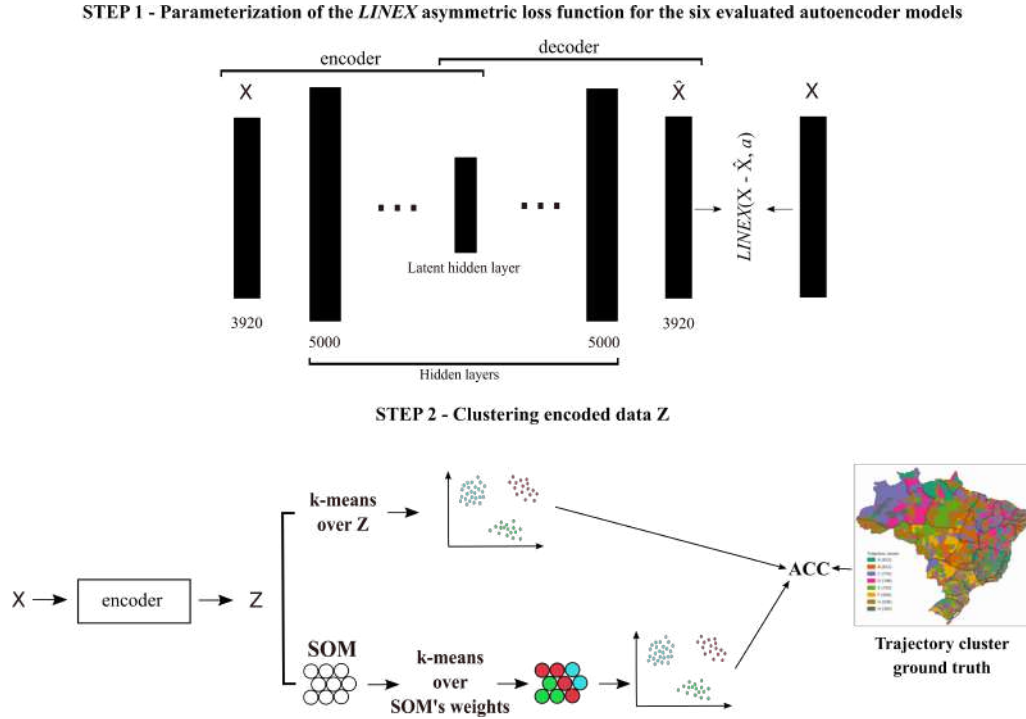


Figure 3. General autoencoder structure used in this paper. Source: elaborated by the authors.

3.2.2. Asymmetric loss function

Asymmetric loss functions are suitable for cases where bias is relevant or a considerable imbalance in data representation, as in our case (very sparse data), or fault detection [Dress et al. 2018]. The most frequently used loss functions for regression are linear and quadratic and have their asymmetric variants [Berk 2011, Liao et al. 2019]. The most relevant characteristics of an asymmetric function are its capability to cope with different error situations and directions. [Gupta et al. 2020] proposed an asymmetric function based on [Huber 1964] that is quadratic when the error is small but is like mean absolute error when the error is larger than a threshold.

Table 1. Encoder layers structure (number of neurons per hidden layer). Source: elaborated by the authors.

ID	Latent*	Encoder layers structure
I	500	3920-5000-3000-2000-1000- 500
II	250	3920-5000-3000-2000-1000-500- 250
III	100	3920-5000-3000-2000-1000-500-250- 100
IV	50	3920-5000-3000-2000-1000-500-250-100- 50
V	25	3920-5000-3000-2000-1000-500-250-100-50- 25
VI	10	3920-5000-3000-2000-1000-500-250-100-50-25- 10

*Number of neurons on the latent layer.

As stated in section 3.1, the standard mean square error loss function and the linear and quadratic asymmetric variant functions failed to guide the learning process of the evaluated autoencoders to reach a convergence curve. Thus, we assessed the linear exponential (LINEX) loss function that rises exponentially on one side of the zero and almost linearly on the other side of the zero [Khatun and Matin 2020, Varian 1975].

The LINEX loss function is given by the Eq. 1, where $\hat{x}_i$ represents the model-based forecast of actual x_i for case i , and $a \neq 0$ is a constant that determines the degree of asymmetry. The direction of the asymmetry can be defined by the signal of a or by change the subtraction $(x_i - \hat{x}_i)$ by $(\hat{x}_i - x_i)$. For $|a| \rightarrow 0$ then the $LINEX(\hat{x}_i) \rightarrow MSE(\hat{x}_i)$, so the $LINEX$ loss function could be thought of as an asymmetric generalization of the mean squared error loss function [Mohammed et al. 2022, Khatun and Matin 2020, Varian 1975].

$$LINEX(\hat{x}_i) = \frac{1}{n} \sum_{i=1}^n \frac{2}{a^2} \left(e^{a(x_i - \hat{x}_i)} - a(x_i - \hat{x}_i) - 1 \right) \quad (1)$$

We evaluated $a \in \{5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 11.0, 12.0\}$ using cross-validation with hold-out method splitting the data into training (80%) and test (20%) datasets. We observed that the autoencoders demonstrated some convergence for $a \geq 5.0$.

3.2.3. Clustering encoded data

To establish a baseline, we clustered the transformed raw data using a k-means algorithm based on joint-trajectories [Genolini et al. 2015] over all 5570 municipalities, 196 variables for 20 years, for $k = 8$ to be able to compare with the results obtained by [Silva et al. 2022].

After defining the a value for the linex loss function, we performed Deep Learning using the entire dataset for each autoencoder model. We conducted clustering on the encoded data using the k-means and the Self-Organizing Map (SOM) [Kohonen 2001] for encoded data reduction and k-means on the SOM's weights. To compare these two methods, we used four clustering validity indices: Silhouette [Rousseeuw 1987], Davies-Bouldin [Davies and Bouldin 1979] using centroids and medoids, and CDbw [Halkidi and Vazirgiannis 2008]. All clustering was conducted considering $k = 8$.

The six deep clustering and the k-means were compared using an accuracy (ACC) measure for the eight clusters (Eq. 2), where b_i represents the ground truth for the municipality i , c_i the cluster obtained by the evaluated clustering method and m a mapping function based on the Hungarian method [Kuhn 1955] to match k_i and c_i . A greater ACC means a good match between our clustering and the one obtained by [Silva et al. 2022].

$$ACC = \max_m \frac{\sum_{i=1}^n \mathbf{1}(b_i = m(c_i))}{n} \quad (2)$$

To check for spatial dependence and regional and intra-regional distinction, we mapped the clustering into the Brazilian municipal geographical map.

3.3. Software

We modeled the deep neural networks using Python and keras framework, clustered using R packages, and used SOMPAK [Kohonen et al. 1996] for the Self-Organizing Map processing. The geographical maps were generated by QGIS version 3.6.

4. Results and discussion

4.1. LINEX parameterization

Fig. 4 shows the mean loss for the train and test dataset for 30 runs (50 epochs each), randomly changing train and test data. We observed these values considering the MSE loss function and for different values for the a parameter for the LINEX loss function. For all loss functions and datasets (train and test), the MSE is almost a horizontal line denoting that the autoencoders did not converge with this loss function. For both training and test datasets, the LINEX loss function with $a = 7$ presents the best result for all autoencoder models. In fact, the performance decreases for $a < 7$ and $a > 7$.

4.2. Clustering encoded data Z

After defining a value for the LINEX loss function parameter a , we presented all datasets to the deep learning process for each evaluated autoencoder model. After that, we encoded the data reducing its dimensionality to the latent layer dimension. Then we proceeded to the clustering using two strategies: a) only k-means over the encoded data; b) and the Self-Organizing Map as an encoded data ordering and reduction and k-means over the SOM's weights after an unsupervised shallow learning process. We evaluated three SOM sizes (8×6 , 10×15 , and 20×15), and we chose the big one because it presented the best quantization error.

Fig. 5 shows the results for the clustering validity indices for both k-means and SOM+kmeans strategies. The Davies-Bouldin presents the same behavior using centroids and medoids as centrality references. Hence, they decay until the smallest values for the autoencoder VI with 10 neurons on the latent layer. The Silhouette index increases while the number of neurons on the latent layer decreases for both clustering methods. The CDbw validity index presents a more erratic behavior for the k-means clustering and a more smooth curve for the SOM 20×15 + k-means approach. All this suggests that the autoencoder would achieve the best data partition with fewer neurons on the latent layer. Still, an inspection of the geographic projection and the accuracy measure shows that the autoencoder with more neurons on the latent hidden layer clustered the Brazilian municipalities more appropriately.

Fig. 6 shows the clustering accuracy when compared with the ground truth (Fig. 2a) for all autoencoder models and clustering algorithms. We observe that clustering using SOM + k-means strategy performed better in most situations, and this suggests that clusters may have a non-convex structure.

Figs. 7 (a) and (c) show the geographic mapping for the k-means partition over the raw data X and for the best clustering considering the ACC as a metric. The k-means clustering over the raw transformed data showed a solid regional spatial dependence, highlighting a clear distinction between the semi-arid region, the states of Amazon, Minas Gerais, and São Paulo, and the Brazilian South (Fig. 7a). The partition using

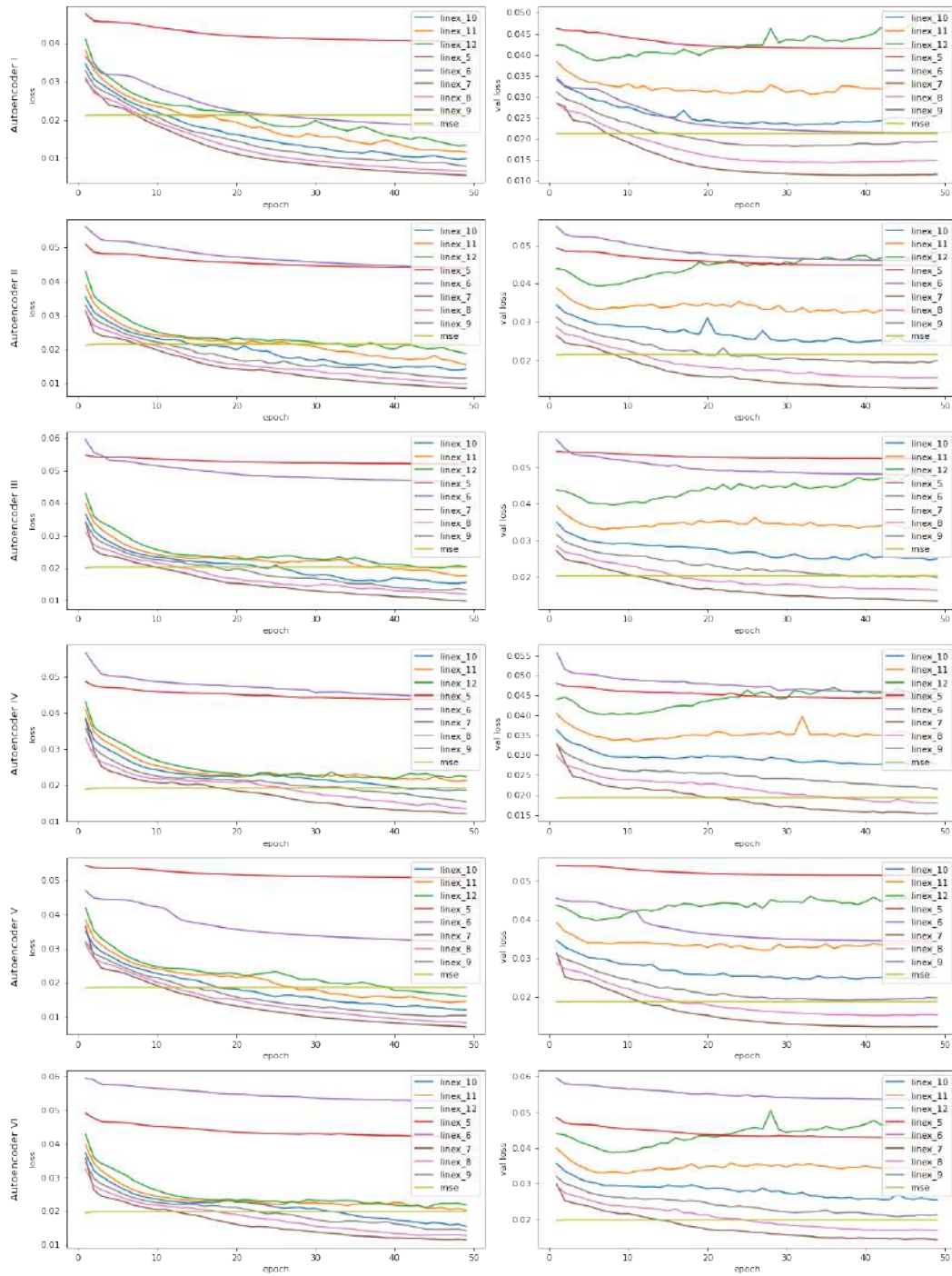


Figure 4. Mean loss values for train and test data considering different loss functions (MSE and LINEX for different α values) considering fifty runs for each autoencoder model. Source: elaborated by the authors.

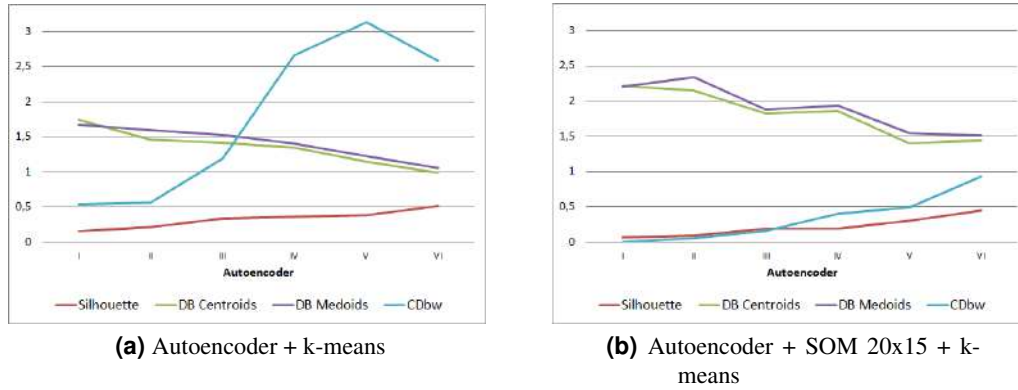


Figure 5. Clustering validity indices for the six autoencoder models considering two clustering strategies (autoencoder + k-means over the encoded data Z , and autoencoder + SOM + k-means over the SOM's weights). Source: elaborated by the authors.

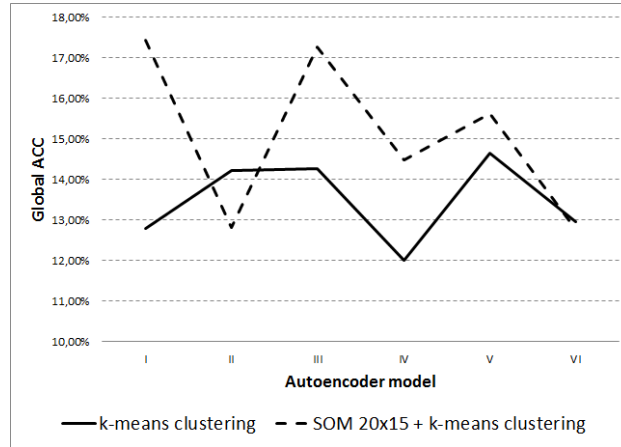


Figure 6. Clustering accuracy (ACC) considering the trajectory cluster from [Silva et al. 2022] as the ground truth. Source: elaborated by the authors.

autoencoder+SOM+k-means also identified regional patterns, separating Minas Gerais from the rest of the country. The Center-West was split into many groups, with Mato Grosso do Sul showing more similarity with the Paraná state.

Figs. 7 (b) and (d) shows intra-regional distinction in the Minas Gerais (MG) Brazilian state for k-means and autoencoder+SOM+k-means strategies. The k-means algorithm did not identify intra-regional patterns in MG and put in one group almost all municipalities of the Center-West and North regions Fig. 7. The clustering based on autoencoder+SOM+k-means identified more intra-regional patterns as observed in Minas Gerais (MG) in Fig. 7d.

5. Conclusions

Performing a clustering analysis over the raw data using a simple k-means algorithm showed that the clusters divide the municipalities showing a huge regional partition but did not identify any intra-regional distinctions. On the other hand, the Deep Clustering

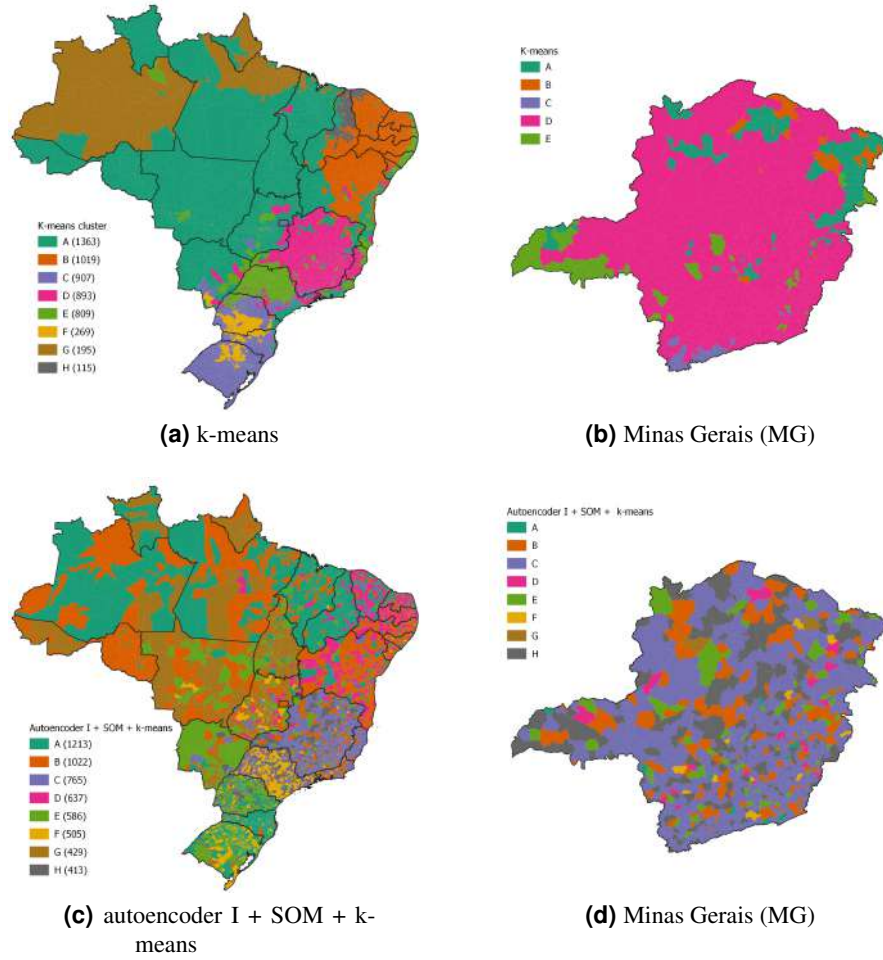


Figure 7. Geographic mapping of the k-means clustering, $k = 8$, applied directly over the transformed raw data X using the `kml3d` function from the `kml` R package (a) and MG (b). Geographic mapping of deep clustering for the autoencoder model I + SOM + k-means (c) and MG (d). In parentheses is the number of observations per cluster. Source: elaborated by the authors.

proceeded in two steps. Dimensionality reduction by autoencoders and clustering of this new data representation using SOM and k-means improved the general accuracy compared to the k-means strategy over the raw data.

Future work should include evaluating larger latent layers and other nonlinear dimensionality reduction techniques such as Isomap [Tenenbaum et al. 2000], and kernel PCA [Müller et al. 2001], exploring other Deep Clustering techniques such as the combination of objective and clustering loss functions as in [Song et al. 2014], the use of multi-view clustering as in [Du et al. 2021], or using variational autoencoders as in [Xu et al. 2020].

References

- Berk, R. (2011). Asymmetric loss functions for forecasting in criminal justice settings. *Journal of Quantitative Criminology*, 27(1):107–123.
- Charte, D., Charte, F., García, S., Jesus, M. J., and Herrera, F. (2018). A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion*, 44:78–96.
- Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2(PAMI-1):224–227.
- Dress, K., Lessmann, S., and Mettenheim, H.-J. (2018). Residual value forecasting using asymmetric cost functions. *International Journal of Forecasting*, 34(4):551–565.
- Du, G., Zhou, L., Yang, Y., Lü, K., and Wang, L. (2021). Deep multiple auto-encoder-based multi-view clustering. *Data Science and Engineering*, 6:323–338. 10.1007/s41019-021-00159-z.
- Falissard, L., Faghreazzi, G., Howard, N., and Falissard, B. (2018). Deep clustering of longitudinal data. *ArXiv*.
- Fatch, P., Masangano, C., Hilger, T., Jordan, I., Mambo, I., Francesca, J., Kamoto, M., Kalimira, A., and Nuppenau, E.-A. (2021). Holistic agricultural diversity index as a measure of agricultural diversity: A cross-sectional study of smallholder farmers in Lilongwe district of Malawi. *Agricultural Systems*, 187:102991.
- Genolini, C., Alacoque, X., Sentenac, M., and Arnaud, C. (2015). kml and kml3d: R packages to cluster longitudinal data. *Journal of Statistical Software*, 65(4):1–34.
- Guo, X., Liu, X., Zhu, E., and Yin, J. (2017). Deep clustering with convolutional autoencoders. *Lecture Notes in Computer Science*, (10635):373–382. 10.1007/978-3-319-70096-0_39.
- Gupta, D., Hazarika, B. B., and Berlin, M. (2020). Robust regularized extreme learning machine with asymmetric huber loss function. *Neural Computing and Applications*, 32:12971–12998.
- Halkidi, M. and Vazirgiannis, M. (2008). A density-based cluster validity approach using multi-representatives. *Pattern Recognition Letters*, 29:773–786.
- Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.
- IBGE (2021). Tabelas 74, 94, 289, 291, 1612, 1613, 3939 e 3940: sistema IBGE de recuperação automática. Available at <https://sidra.ibge.gov.br> (2021/06/15).
- Khatun, N. and Matin, M. A. (2020). A study on linex loss function with different estimating methods. *Open Journal of Statistics*, 10:52–63.
- Kohonen, T. (2001). *Self-Organizing Maps*. Berlin: Springer.
- Kohonen, T., Hynninen, J., Kangas, J., and Laaksonen, J. (1996). SOM PAK: The self-organizing map program package. Technical report, Espoo: Finland.
- Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.
- Liao, L., Park, C., , and Choi, H. (2019). Penalized expectile regression: an alternative to penalized quantile regression. *Annals of the Institute of Statistical Mathematics*, 71(2):409–438.
- Min, E., Guo, X., Liu, Q., Zhang, G., Cui, J., and Long, J. (2018). A survey of clustering with deep learning: From the perspective of network architecture. *IEEE Access*, 6:39501–39514. 10.1109/ACCESS.2018.2855437.
- Mohammed, M., Alshanbari, H. M., and El-Bagoury, A.-A. H. (2022). Application of the linex loss function with a fundamental derivation of liu estimator. *Computational Intelligence and Neuroscience*, (2307911):–. Artificial Intelligence and Machine Learning-Driven Decision-Making.
- Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., and Schölkopf, B. (2001). An introduction to kernel-based learning algorithms. *IEEE TRANSACTIONS ON NEURAL NETWORKS*, 12(2):181–201.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Computational and Applied Mathematics*, 20:53–65.
- Sales, C. and Rodrigues, R. (2019). Espaço rural brasileiro: diversificação e peculiaridades. *Revista Espinhaço*, 8(1):54–65.
- Silva, M. A. S. d., Matos, L. N., Santos, F. E. d. O., Dompieri, M. H. G., and Moura, F. R. d. (2022). Tracking the connection between brazilian agricultural diversity and native vegetation change by a machine learning approach. *IEEE Latin America Transactions*, 20(11):2371–2380.
- Song, C., Y, Y. H., Liu, F., Wang, Z., and Wang, L. (2014). Deep auto-encoder based clustering. *Intelligent Data Analysis*, 18(6):S65–S76. 10.3233/IDA-140709.
- Teixeira, M. and Ribeiro, S. (2020). Agricultura e paisagens sustentáveis: a diversidade produtiva do setor agrícola de Minas Gerais, Brasil. *Sustainability in Debate*, 11(2):29–41.
- Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323.
- Varian, H. R. (1975). A bayesian approach to real estate assessment. *Studies in Bayesian Econometric and Statistics in Honor of Leonard J. Savage*, 5:195–208.
- Xu, C., Dai, Y., Lin, R., and Wang, S. (2020). Deep clustering by maximizing mutual information in variational auto-encoder. *Knowledge-Based Systems*, 205(106260). 10.1016/j.knosys.2020.106260.

A New Big Data Architecture for Efficient Processing of Spatiotemporal Data using Machine Learning

Sávio S. Teles de Oliveira¹, Vagner J. do Sacramento Rodrigues¹,
Wellington S. Martins¹, Anderson da Silva Soares¹

Federal University of Goiás (UFG)
Alameda Palmeiras, Quadra D, Câmpus Samambaia
131 - CEP 74001-970 - Goiânia - GO - Brazil

{savio.oliveira, vagner, wellington, anderson}@inf.ufg.br

Abstract. *The volume of spatiotemporal data is increasing in the era of Big Data and has brought several challenges in managing and processing these data. Many big data platforms have been created to address these challenges, but none can outperform the others in all scenarios on spatiotemporal queries. This paper presents a scalable architecture for efficient spatiotemporal data processing. This architecture allows a couple of several big data platforms and automatically choose on the fly, using machine learning, the best big data platform to process each spatiotemporal query.*

1. Introduction

Spatiotemporal data comprises spatial and temporal representations, and it has attracted great interest in the world, becoming critical in several areas such as health [Roy et al. 2021] and social networking [Parimala et al. 2021]. Large amounts of spatiotemporal data are being generated and captured through systems that record sequential observations of remote sensing, mobility, wearable devices, and social media. This large volume of data brings unique challenges in storing, managing, and processing these data, leading these challenges to the Big Spatiotemporal Data [Yang et al. 2020].

Several big data platforms have been proposed in the literature [Wang et al. 2020] to handle spatiotemporal data, such as Hadoop [Hadoop 2022], Spark [Spark 2022], Elasticsearch [Elasticsearch 2022] and SciDB [Paradigm4 2022]. Some papers [Zhang et al. 2018, Karim et al. 2018] show that no big data platform performs better than the others in all scenarios with spatiotemporal filters. It is up to the solution developer to choose in advance which platform they want to use for each scenario. This choice is a complicated task to be made dynamically because each big data platform has characteristics that allow it to perform best for specific spatial and temporal filtering scenarios. In addition, the platforms are constantly evolving and releasing new versions, and optimizations may emerge that overlap with competing platforms.

A new processing engine, SmarT, is presented in [de Oliveira et al. 2020, de Oliveira et al. 2021] for filtering and retrieving spatiotemporal data in big data using machine learning algorithms. SmarT seeks to choose, in real time, the best system to process the request. However, SmarT has some limitations: i) it is not extensible for any machine learning algorithm, ii) it doesn't have a complete and scalable architecture for managing and storing spatiotemporal data, iii) it cannot index the spatiotemporal data, and iv) it is not scalable on processing spatiotemporal queries.

Our work presents a new scalable architecture for efficiently processing of spatiotemporal data. Furthermore, it improves the SmarT with a new flexible architecture to be able to include new machine learning algorithms. Therefore, we evaluate many machine learning methods to determine which one we should choose to predict the big data platform to process each query. The architecture proposed in this work is an efficient, scalable, and robust solution for processing large volumes of spatiotemporal data.

The remainder of the paper is organized as follows. Section 2 describes the strategies found in the literature for processing of spatiotemporal data. Section 3 presents the details of our architecture, and Section 4 describes the experiment’s methodology and results. Section 5 presents the conclusions of this work and a description of future works.

2. Related Works

There are many attempts to processing spatiotemporal data in database systems. Due to the high computational cost and the large volume of spatiotemporal data available, some works in big data [de Oliveira et al. 2019, Chi et al. 2016] introduced parallel and distributed solutions for processing spatiotemporal data. The MapReduce programming model has caused a revolution in the processing and management of spatiotemporal data because it has facilitated the development of works for processing spatiotemporal data [de Assis et al. 2017, Hamdi et al. 2022].

The Apache Hadoop [Hadoop 2022] and Apache Spark [Spark 2022] rapidly gained popularity among the scientific community due to the efficient spatiotemporal data processing. Some spatiotemporal extensions for Spark, like SpatialSpark [You et al. 2015], and Hadoop, like SpatialHadoop [Eldawy and Mokbel 2015], emerged to provide efficient spatial operations. Popular array databases like Rastaman [Baumann et al. 1997] and SciDB [Brown 2010] show promising performance gains compared to traditional methods [Xu et al. 2020]. These databases have been used in several works [Lu et al. 2016, Camara et al. 2016] for efficient processing of spatiotemporal data. Beyond the SciDB, Elasticsearch is also recommended for spatiotemporal data processing [Shrivastava 2020].

Some works [Doan et al. 2016, Zhang et al. 2018] show that no big data platform is more efficient than the others in all spatiotemporal filter scenarios. In [de Oliveira et al. 2020, de Oliveira et al. 2021], the authors introduced a new query engine (SmarT) that chooses, in real time, the best big data platform to process the spatiotemporal queries. However, SmarT is not scalable and cannot handle highly concurrent queries. An efficient and scalable architecture for spatiotemporal data ingestion and processing becomes necessary for multiple big data platforms and simultaneous queries.

3. Exploring Big Data Architectures for Spatiotemporal Data Processing

In order to manage a large volume of spatiotemporal data, this work presents a new distributed architecture comprising an indexing and storing system, and a module for processing spatiotemporal queries. This architecture, shown in figure 1, can be coupled with any big data platform for processing spatiotemporal data, such as SciDB, Apache Spark, and Elasticsearch. Our architecture uses machine learning algorithms to choose, in real time, which big data platform will process each query. In other words, the developer does not need to select, in advance, which big data platform will process each type of query.

This architecture supports the insertion and indexing of new data (shown in section 3.1) and can also process queries using spatiotemporal filters (details in section 3.2). The architecture keeps a distributed message queuing to facilitate the communication and coordination between the processing jobs. The distributed queues can significantly improve performance, reliability, and scalability when processing spatiotemporal data.

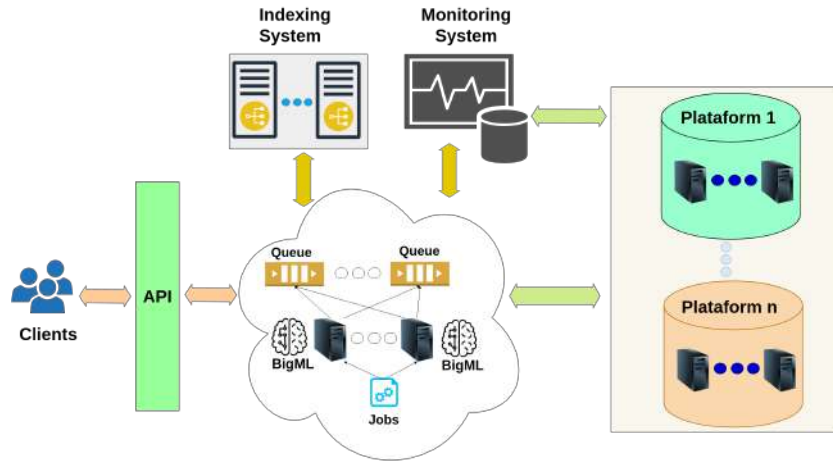


Figure 1. Architecture for efficient spatiotemporal data processing.

3.1. Inserting and Indexing Operations on Spatiotemporal Data

On insertion, the API has as input a set of spatiotemporal objects, where each object contains the following properties: i) geographic location in OGC Well-Known Text (WKT) format with a precision of at most 15 decimal places; ii) *timestamp* in UNIX format in seconds¹; iii) object metadata in JSON format. With these three properties, it is possible to define each spatiotemporal object and also allow developers to add object metadata according to their business rules.

The architecture inserts the spatiotemporal data in a distributed way using *jobs* for data processing. Each job has a specific role in the data pipeline. The first job, *J1*, gets the data from the API and sends it to the queue. The second job, *J2*, gets the objects in the queue and sends each object's spatial location and timestamp to the Indexing System to index spatially and temporally. As seen in Figure 1, the Indexing System has an exclusive cluster to index these objects in a distributed way. Job *J2* keeps a list of servers in the Indexing System cluster and sends each object to a server to be indexed. This server is chosen randomly from the list.

Each server has a BKD-Tree index (Block K-D tree) [Procopiuc et al. 2003] to keep the spatiotemporal data indexed efficiently. BKD-Tree is designed to be I/O efficient by keeping most of the data structures in disk blocks, while a small binary tree index structure is stored in RAM. Our work uses Elasticsearch as a distributed Indexing System to coordinate the local indexing of each BKD-Tree. Elasticsearch is a high-performing solution in indexing spatiotemporal [Wang et al. 2019] data.

¹https://pubs.opengroup.org/onlinepubs/9699919799/basedefs/V1_chap04.html#tag_04_16

Once indexed, the data pipeline moves on to the next step of storing the object in each big data platform. A message is published on a queue with the object's data. There are n jobs listening to this queue to store the spatiotemporal data across platforms, where n is the number of big data platforms. Each of these n jobs is responsible for storing the new data on a specific platform and knows how to build the object to store on that platform. This architecture facilitates the addition of new platforms, just creating a particular job for each new platform.

This flexible architecture allows us to easily add new steps to the data pipeline. That happens because the distributed queues keep a transparent communication between the producers and consumers in each pipeline step. In this architecture, we can have multiple jobs replicas consuming the data from the queues because the operations are independent, and insertions or indexing operations can be done in parallel. Therefore, the architecture is scalable and elastic due to the ability to add new data processing nodes simply by increasing the number of machines and replicas of the processing jobs.

3.2. Using Machine Learning for Efficient Spatiotemporal Query Processing

Our work proposes the BigML that aims to minimize the response time of spatiotemporal queries by choosing the big data platform with the lower response time predicted by the machine learning algorithm. Since query response time is a continuous variable, the problem of minimizing it is a task of approximating a mapping function f from input variables X to a continuous output variable y (query response time).

The client sends the spatiotemporal query parameters and gets the results through the API. One job S reads the query parameters from the API and becomes the query coordinator. So, this job will decide how to process the query, format the responses, and send the result to the client. Using the architecture proposed in figure 1, it is possible to scale the jobs quickly, which allows the processing of many queries in parallel. The job uses the BigML to choose which big data platform will process the query. BigML has the following input: (i) the spatial and temporal query constraints; ii) the cluster metrics acquired from the monitoring system, and iii) the approximate number of matches for the spatiotemporal query obtained in the indexing system. So, the job is responsible for collecting these inputs to send to BigML.

First, the job S has to get cluster metrics on the monitoring system. The job sends a request to the monitoring system that returns the current cluster metrics. Next, the job S obtains the approximate count of matches for the query in the indexing system. The job then queries the BigML that returns which big data platform must process the query. Afterward, S format the query for the big data platform chosen by BigML and request it to process the query. Then, the big data platform runs the query and returns the results for the job. Finally, the job format the output and sends it to the client.

3.2.1. Training the BigML models

The machine learning model training is triggered manually using data collected over time from client requests. The first training dataset is generated using simulated user requests. The training happens periodically, and at the end of it, the output model is updated on BigML, which uses it to predict the response time of each big data platform to process

users' queries. BigML has to train and keeps a model for each *big data* platform, where the dependent variable represents the response time of the *big data* platform, and the query parameters are the (independent) explanatory variables. We can denote the response variable by y and the set of explanatory variables by $x_1, x_2, \dots, x_p$, where p indicates the number of explanatory variables. The relationship between y and $x_1, x_2, \dots, x_p$ is approximated by the regression model of the equation 1:

$$y = f(x_1, x_2, \dots, x_p) + \varepsilon \quad (1)$$

with ε being a random error representing the discrepancy in the approximation. Thus, the function $f(x_1, x_2, \dots, x_p)$ describes the relationship between y and $x_1, x_2, \dots, x_p$.

There are many regression models [Chatterjee and Hadi 2015] that can be used to estimate the response time of each big data platform. The BigML solution proposed in this paper is flexible to integrate any machine learning model that meets the following requirements: algorithm inputs are pairs of training examples $(\vec{x}_1, y_1), (\vec{x}_2, y_2), (\vec{x}_p, y_p)$, where $\vec{x}$ is a feature vector representing the explanatory variables and y is the query response time. Being aware that the response time is the response variable y , then for each training input pair $(\vec{x}_i, y_i)$, the response variable y_i is defined by the equation 2 and the feature vector $\vec{x}_i$ by the equation 3:

$$y_i = response\_time_i \quad (2)$$

$$\begin{aligned} \vec{x}_i = (&area_i, time\_interval_i, count_i, swap\_free_i, mem\_free_i, \\ &load\_one_i, load\_five_i, load\_fifteen_i, \\ &cpu\_user_i, cpu\_system_i, bytes\_in_i, bytes\_out_i) \end{aligned} \quad (3)$$

The explanatory variables *area* and *time_interval* come from the spatial and temporal query filters, and *count* is generated by the Indexing System, which returns the estimated number of the query results. By combining the *area*, *time_interval* and *count* variables, it is possible to estimate, more accurately, the cost of processing the query. The other metrics are obtained from the cluster monitoring system and are important because they impact the performance of a big data platform when running a spatiotemporal query. The cluster metrics are collected to represent the state of the cluster at the time of the query. Several other variables can be coupled to the BigML architecture, but this work explores the abovementioned variables.

4. Experiments

This section evaluates the architecture proposed in this paper for processing a large volume of spatiotemporal data. The section 4.2 measures the architecture when ingesting and storing a large volume of data per second. The section 4.3 evaluates the architecture performance when processing spatiotemporal queries using the BigML search engine to choose, in real time, between the big data platforms Elasticsearch, Apache Spark, and SciDB, the best one to run each query.

4.1. Experimental Environment

The architecture integrates three big data platforms in our work: i) Elasticsearch version 7.0.1, ii) SciDB version 18.3, and iii) Apache Spark version 2.3.2. The platforms were

deployed on eight servers in the Amazon EC2² Cloud Computing environment. Each server has 3.1 GHz Intel Xeon Platinum 8175 processors with 16 vCPUs, 128 GB of RAM and 600 GB of SSD. The computers were connected by a 10 Gbit/second Ethernet network with dedicated bandwidth of 3500 Mbps.

More than 7.2 billion observations related to 16.7 million pixels of an area of 800.824 km^2 covering a piece of the Midwest and Southeast of Brazil have been stored on every big data platform. Each pixel has a time series with 435 observations between 2000 and 2019 extracted from NASA's MOD13Q1 product³ (National Aeronautics and Space Administration), with a periodicity of 16 days between observations and a spatial resolution of 250 meters.

4.2. Evaluation of Data Ingestion and Distributed Storage

This section aims to evaluate the architecture's performance in collecting, transforming, and ingesting a large volume of spatiotemporal data. In this evaluation, we use Apache Kafka on version 2.0.0 as the message exchange system and Spark Streaming, version 2.3.2, as the engine for creating the data processing jobs. The data is replicated for indexing and storing on the big data platforms Elasticsearch, SciDB, and Apache Spark.

Three servers were used in the Amazon EC2 Cloud Computing environment to receive, transform, and insert the datasets. Each server has an Intel Xeon Platinum 8175 processor with 16 vCPUs, 128 gigabytes of RAM, and 600 gigabytes of storage. The machines were connected to a network with a dedicated bandwidth of over 3000 Mbps. The input data is sent as messages to Kafka containing the temporal observations with the pixel band values and vegetation indices. One machine was used to send the query requests to the cluster and collect the metrics. This machine has a 2.4 GHz Intel Xeon E5-2676 v3 processor with 8 cores, 32 GB of RAM, and 1TB of SSD.

4.2.1. Results and Discussions

Experiments were carried out to measure the number of messages and the volume of data that the architecture can index and store. Figure 2(a) shows the number of messages processed per second during 200 seconds. It processed up to 6.2 million messages per second, keeping an average of approximately 5.8 million messages processed per second over time. The standard deviation of roughly 630,000 messages processed per second shows that the platform maintained a spatiotemporal data processing rate over time.

Figure 2(b) presents the volume of messages in MB processed per second in 200 seconds by the architecture. Each message has approximately 40 bytes because they were compressed using the lz4 algorithm before being sent to Apache Kafka to increase the throughput. The proposed architecture processed up to 164 MB per second, with an average of 148 MB/second. The consumption in MB of messages per second remained constant, with a standard deviation of 15 MB/second.

Table 1 shows how long takes to store and index the 7.2 billion spatiotemporal data in Elasticsearch, Spark, and SciDB. Spark had the shortest time (780 seconds) to store all

²<https://aws.amazon.com/pt/ec2/>.

³This dataset is available by LAPIG (Image Processing and Geoprocessing Laboratory)

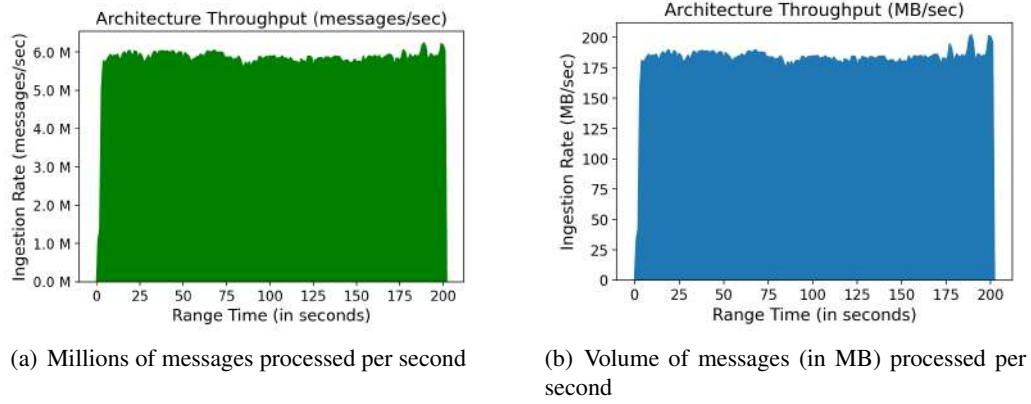


Figure 2. Evaluation of the architecture in processing the spatiotemporal data.

the data because it compresses with Snappy and encodes the data in parquet format, and also it does not build an indexing system on top of the data. Apache Spark had insertion throughput almost 39 times higher than Elasticsearch and 43 times higher than SciDB. In addition, Spark consumed only 12.7 GB of disk space to index and store the data in parquet format, which is 21 times lower than Elasticsearch and 20 times less than SciDB.

	Apache Spark	Elasticsearch	SciDB
Total time (seconds)	780	30 360	33 534
Dataset size (in GB)	12,7	273	262

Table 1. Time to store the dataset in each big data platform.

Elasticsearch required 30,360 seconds to store and build spatial and temporal indexes. This indexing has a high computational cost because each pixel has to be indexed spatially (using its latitude and longitude) and temporally, using the date of each observation. The spatial indexing has less weight than the temporal indexing because the 16,777,216 *pixels* are indexed only once. But, the temporal indexing significantly impacted the Elasticsearch throughput, since it had to index all 7.2 billion observations.

SciDB stored and indexed the data in 33534 seconds. It was worse than Elasticsearch because it indexed spatially and temporally each of the 7.2 billion observations. In its dimension-based indexing system, each observation's latitude, longitude, and date are stored as an array with three dimensions. Each of these three dimensions must be indexed to allow spatiotemporal filters to be applied to queries. SciDB consumed 9 GB more disk space than Elasticsearch to store the data. Elasticsearch and SciDB have a larger storage footprint than Spark because they lack efficient formats for storing spatiotemporal data.

4.3. Evaluation of Query Processing with BigML

This section evaluates the BigML when deciding which big data platform is better for each query, analyzing the following machine learning models. First, the Generalized Linear Model (GLM) [Nelder and Wedderburn 1972] setting *family* = *gaussian* and *alpha* = 0. Second, the Distributed Random Forest (DRF) [Cutler et al. 2012] using *score_tree_interval* = 5 and *distribution* = *gaussian*. Third, the Gradient

Boosting Machine (GBM) [Friedman 2001] with *ntrees* = 68, *max_depth* = 11, and *min_rows* = 1. Fourth the XGBoost [Chen and Guestrin 2016] with the parameters *ntrees* = 93, *max_depth* = 12, and *min_rows* = 5. Finally, the Deep Learning [LeCun et al. 2015] model using *epochs* = 400 and *activation* = *RectifierWithDropout*. We have used the open source library H2O ⁴ in version 3.36. The experiments use the K-fold cross-validation method, with k=5, using the default values for the remaining H2O algorithms' parameters.

In all, 1983 spatial and temporal constraints were generated from: (i) 247 spatial filters extracted from the region of the state of Goiás in Brazil, ii) 7 temporal filters in the interval from 2016 to 2019 and iii) 1729 (247 * 7) spatiotemporal filters combining the 247 spatial and 7 temporal filters. The queries return from 0 to 2 billion results using these spatial and temporal filters. Each query ran ten times on each of the three big data platforms. We collected cluster metrics ($\vec{x}_i$) and the query response time (y_i) as input for training each machine learning algorithm. We extracted 75% for the training set and 25% for the test set.

SciDB and Apache Spark were installed with the default settings, without any optimization or customization. Regarding Elasticsearch, the JVM memory default was increased from 1 GB to 48 GB, as queries were causing memory overflow during testing. The tests on Spark ran using the Hadoop ecosystem version 3.1.1, with Apache Yarn scheduling the queries and managing the cluster resources consumed by Spark. The data were stored in HDFS on parquet format with snappy ⁵ compression.

4.3.1. Results and Discussion

This section presents the results of evaluating BigML in predicting the big data platform to process each query. BigML predicts the query response time for each big data platform tied to the architecture (e.g., Elasticsearch, SciDB, and Spark). BigML chooses the platform with the lowest response time predicted. The architecture is flexible to use different machine learning algorithms for this prediction, and the table 2 presents a comparison between the five algorithms evaluated in this work.

Table 2 compares the accuracy and RMSE of the machine learning models. The highest accuracy was identified in the GBM and XGBoost models, 92.1% and 90.9%, respectively. Some studies show that these boosting methods perform well with tabular input data, which is our input format, for two main reasons [Nielsen 2016]: (i) ability to create a rich representation of the data by approximating complex relationship functions between the data, including interactions in high-dimensional spaces; ii) versatility to adapt to different sets of explanatory variables by automatically performing feature selection.

The XGBoost is the model used by SmarT [de Oliveira et al. 2021], and table 2 shows that GBM has a better accuracy in this task. Our proposed solution, BigML, is flexible in using any machine learning method and the remaining evaluation in this Section considers that BigML uses the GBM algorithm. BigML achieved an overall accuracy of 92.1%, in choosing the best platform compared to SmarT with 90.9%. This result is

⁴<https://www.h2o.ai/>

⁵<https://github.com/google/snappy>

satisfactory because the BigML using the GBM algorithm achieved a high accuracy that reduces the response time from spatiotemporal queries by choosing the Big Data platform to run each query.

	Overall Accuracy	RMSE_{Elastic}	RMSE_{Spark}	RMSE_{SciDB}
GLM	31.8%	9378	1752	434
DRF	75.2%	1373	242	123
GBM	92.1%	334	165	101
XGBoost (SmarT)	90.9%	369	164	99
Deep Learning	77.7%	1218	295	138

Table 2. Comparison of the accuracy and RMSE of machine learning algorithms.

Table 2 shows that the RMSE_{Elastic} (concerning to Elasticsearch) is the highest among the three big data platforms. This result can be explained because Elasticsearch has a huge variation in response time as the size of query results increases. This causes the task of predicting the query response time more difficult. Elasticsearch has a very efficient internal indexing mechanism, but its performance depends on the size of the results. When it gets larger than a certain limit ⁶, the response time increases substantially [Thacker et al. 2016].

When the number of results increases, Spark and SciDB perform better. Spark stores the data in Parquet format and can retrieve a larger dataset faster than Elasticsearch. SciDB has an efficient internal system for retrieving spatiotemporal data from disk. It can perform better than Spark for many results because it exploits its efficient indexing and retrieval system for multidimensional data [Zhang et al. 2016]. Spark, by contrast, has to scan a lot of Parquet files to filter the results. When the dataset is larger than the memory size, the intermediate results need to be serialized on Spark and stored on secondary storage devices (disk or SSD) or recalculated

Table 3 compares the mean response time for the spatiotemporal queries with BigML, using GBM algorithm, when all big data platforms are available or just two of them. As seen in Table 3, using BigML, the query response time was 490ms versus 482ms if it had always chosen the best platform (“Best” column). Even when BigML does not select the best platform, there is a slight difference in response time if it has always chosen the best one for each query. It is recommended to have all three platforms to have the best possible performance. But if it is not possible, table 3 shows that keeping Spark and SciDB is the best choice. SciDB proved to perform better than the others when choosing only one big data platform. Table 3 shows that BigML had a high accuracy when Elasticsearch was deployed with Spark or SciDB because it could identify the features that modeled the high response time variance of Elasticsearch.

5. Conclusion

The large volume of spatiotemporal data, with high dimensionality being generated in large quantities at all times poses the problem of processing it in the context of Big Data. The efficient filtering and retrieval of a large volume of data are one of the biggest open

⁶This limit depends heavily on the Elasticsearch and the cluster configuration.

	Best	BigML	Spark	Elasticsearch	SciDB
All platforms	482	490	822	6291	605
Elasticsearch + SciDB	533	533	∅	6291	605
Elasticsearch + Spark	796	796	822	6291	∅
Spark + SciDB	513	522	822	∅	605

Table 3. Comparison of the average response time (ms).

challenges in the area. This work proposes a new architecture to process spatiotemporal data efficiently, ingesting and processing a large volume of data per second.

The architecture has a search engine, BigML, that can efficiently process spatiotemporal data, and it is flexible to integrate any machine learning algorithm. In this work, we evaluated five machine learning algorithms, where GBM had the best accuracy with 92.1% against 90.9% from SmarT [de Oliveira et al. 2021] in choosing the best platform among Apache Spark, Elasticsearch, and SciDB to process a spatiotemporal query. The response time decreases 23% using BigML than the best big data platform evaluated (e.g., SciDB). Furthermore, SmarT’s architecture is not scalable and cannot process many parallel spatiotemporal parallel queries, which is mandatory in Spatial Big Data. The architecture proposed in this paper can handle many concurrent queries efficiently using distributed jobs to process each spatiotemporal query.

One disadvantage of the architecture proposed in this work is that data must be replicated among the big data platforms since there is no efficient bus for data storage and retrieval. In future work, we plan to create an efficient data bus that integrates the various big data platforms. Lastly, we also intend to evaluate other machine learning models and the impact of other variables, such as internal metrics for each big data platform.

References

- Baumann, P., Furtado, P., Ritsch, R., and Widmann, N. (1997). The rasdaman approach to multidimensional database management. In *Symposium on Applied Computing: Proceedings of the 1997 ACM symposium on Applied computing*, volume 1997, pages 166–173.
- Brown, P. G. (2010). Overview of scidb: large scale array storage, processing and analysis. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, pages 963–968. ACM.
- Camara, G., Assis, L. F., Ribeiro, G., Ferreira, K. R., Llapa, E., and Vinhas, L. (2016). Big earth observation data analytics: matching requirements to system architectures. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data*, pages 1–6. ACM.
- Chatterjee, S. and Hadi, A. S. (2015). *Regression analysis by example*. John Wiley & Sons.
- Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM.

- Chi, M., Plaza, A., Benediktsson, J. A., Sun, Z., Shen, J., and Zhu, Y. (2016). Big data for remote sensing: Challenges and opportunities. *Proceedings of the IEEE*, 104(11):2207–2219.
- Cutler, A., Cutler, D. R., and Stevens, J. R. (2012). Random forests. In *Ensemble machine learning*, pages 157–175. Springer.
- de Assis, L. F. F. G., de Queiroz, G. R., Ferreira, K. R., Vinhas, L., Llapa, E., Sanchez, A. I., Maus, V., and Câmara, G. (2017). Big data streaming for remote sensing time series analytics using mapreduce. *Revista Brasileira de Cartografia*, 69(5).
- de Oliveira, S. S. T., Martins, W. S., Sacramento, V., Bueno, E., Cardoso, M., and Pascoal, L. (2019). A parallel and distributed approach to the analysis of time series on remote sensing big data. *Journal of Information and Data Management*, 10(1):16–34.
- de Oliveira, S. S. T., Rodrigues, V. J. S., and Martins, W. S. (2021). Smart: Machine learning approach for efficient filtering and retrieval of spatial and temporal data in big data. *Journal of Information and Data Management*, 12(3).
- de Oliveira, S. S. T., Vagner, J., Martins, W. S., Palmeiras, A., Quadra, D., and Samambai, C. (2020). Smart: Uso de aprendizado de máquina para filtragem e recuperação eficiente de dados espaciais e temporais em big data. In *SBBD*, pages 85–96.
- Doan, K., Oloso, A. O., Kuo, K.-S., Clune, T. L., Yu, H., Nelson, B., and Zhang, J. (2016). Evaluating the impact of data placement to spark and scidb with an earth science use case. In *2016 IEEE International Conference on Big Data (Big Data)*, pages 341–346. IEEE.
- Elasticsearch (2022). Elasticsearch. <https://www.elastic.co>.
- Eldawy, A. and Mokbel, M. F. (2015). Spatialhadoop: A mapreduce framework for spatial data. In *2015 IEEE 31st international conference on Data Engineering*, pages 1352–1363. IEEE.
- Friedman, J. H. (2001). Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.
- Hadoop (2022). Apache hadoop. <https://hadoop.apache.org>.
- Hamdi, A., Shaban, K., Erradi, A., Mohamed, A., Rumi, S. K., and Salim, F. D. (2022). Spatiotemporal data mining: a survey on challenges and open problems. *Artificial Intelligence Review*, 55(2):1441–1488.
- Karim, S., Soomro, T. R., and Burney, S. A. (2018). Spatiotemporal aspects of big data. *Applied Computer Systems*, 23(2):90–100.
- LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.
- Lu, M., Pebesma, E., Sanchez, A., and Verbesselt, J. (2016). Spatio-temporal change detection from multidimensional arrays: Detecting deforestation from modis time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 117:227–236.
- Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

- Nielsen, D. (2016). Tree boosting with xgboost-why does xgboost win" every" machine learning competition? Master's thesis, NTNU.
- Paradigm4 (2022). Scidb. <https://www.paradigm4.com>.
- Parimala, M., Swarna Priya, R., Praveen Kumar Reddy, M., Lal Chowdhary, C., Kumar Poluru, R., and Khan, S. (2021). Spatiotemporal-based sentiment analysis on tweets for risk assessment of event using deep learning approach. *Software: Practice and Experience*, 51(3):550–570.
- Procopiuc, O., Agarwal, P. K., Arge, L., and Vitter, J. S. (2003). Bkd-tree: A dynamic scalable kd-tree. In *International Symposium on Spatial and Temporal Databases*, pages 46–65. Springer.
- Roy, S., Bhunia, G. S., and Shit, P. K. (2021). Spatial prediction of covid-19 epidemic using arima techniques in india. *Modeling earth systems and environment*, 7(2):1385–1391.
- Shrivastava, S. (2020). A review of spatial big data platforms, opportunities, and challenges. *IETE Journal of Education*, 61(2):80–89.
- Spark (2022). Apache spark. <https://spark.apache.org>.
- Thacker, U., Pandey, M., and Rautaray, S. S. (2016). Performance of elasticsearch in cloud environment with ngram and non-ngram indexing. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)*, pages 3624–3628. IEEE.
- Wang, F., Li, M., Mei, Y., and Li, W. (2020). Time series data mining: A case study with big data analytics approach. *IEEE Access*, 8:14322–14328.
- Wang, S., Zhong, Y., and Wang, E. (2019). An integrated gis platform architecture for spatiotemporal big data. *Future Generation Computer Systems*, 94:160–172.
- Xu, M., Zhao, L., Yang, R., Yang, J., Sha, D., and Yang, C. (2020). Integrating memory-mapping and n-dimensional hash function for fast and efficient grid-based climate data query. *Annals of GIS*, pages 1–13.
- Yang, C., Clarke, K., Shekhar, S., and Tao, C. V. (2020). Big spatiotemporal data analytics: A research and innovation frontier.
- You, S., Zhang, J., and Gruenwald, L. (2015). Large-scale spatial join query processing in cloud. In *2015 31st IEEE international conference on data engineering workshops*, pages 34–41. IEEE.
- Zhang, X., Khanal, U., Zhao, X., and Ficklin, S. (2016). Understanding software platforms for in-memory scientific data analysis: A case study of the spark system. In *2016 IEEE 22nd International Conference on Parallel and Distributed Systems (ICPADS)*, pages 1135–1144. IEEE.
- Zhang, X., Khanal, U., Zhao, X., and Ficklin, S. (2018). Making sense of performance in in-memory computing frameworks for scientific data analysis: A case study of the spark system. *Journal of Parallel and Distributed Computing*, 120:369–382.

A Survey on Persistence Strategies for Semantically Enriched Trajectories

Flaris Roland Feller¹, Ronaldo dos Santos Mello¹

¹Depto. de Informática e Estatística – Univ. Federal de Santa Catarina (UFSC)
Caixa Postal 5040 – 88.040-970 – Trindade – Florianópolis – SC – Brasil

flaris.feller@posgrad.ufsc.br, r.mello@ufsc.br

Abstract. *Mobile object databases have been studied for at least 20 years. However, with the emergence of Big Data, new challenges for MODs were introduced, such as the ability to manage a large volume of data, support for low-latency spatiotemporal queries, the need for high insert throughput, and association from complex and heterogeneous data to raw trajectories. This last challenge gives rise to the concept of semantically enriched trajectories, i.e., trajectories with one or more associated semantic contexts, such as the speed of the movement, the points of interest on which the trajectory passes and the weather condition during the movement. This work presents a survey of recent proposals for the persistence of semantically enriched trajectories in databases using different technologies such as NoSQL, relational and triplestore. The characteristics raised and analyzed are the implementation project, taking into account the storage strategy of these trajectories, the database platform used and the considered data models. For the best of our knowledge, there is no survey in the literature with focus on persistence of semantic trajectories.*

1. Introduction

Due to the vast use of GPS-enabled devices such as vehicle navigation systems, smartphones and wearable devices, recording position data has become an easy task and large amounts of data are collected every day. In response to this, trajectory data management research, or *Mobile Object Database* (MOD), has been active for the past 20 years.

A trajectory describes the movement of an entity, for example a person, a vehicle or an animal, over time. At a lower level of abstraction, it is a sequence of positions in (x,y) space with a *timestamp* t , also known as *raw trajectory*, corresponding to the way data is written by the devices [Güting and Schneider 2005]. An example of a raw trajectory is shown in Figure 1 (a).

An important consideration to this context is to attach meaning to the trajectory as a whole or parts of it to discover interesting phenomena. For example, for a tourist, instead of identifying which geographic coordinate he/she is at, we would like to know if he/she is visiting a museum or having dinner at a restaurant. We may also be interested in knowing if this tourist is traveling on foot, by bicycle or using a bus [Güting et al. 2015].

In recent years, the management of mobility data has greatly expanded. In particular, a promising new approach has been developed to provide applications with richer and more meaningful knowledge about motion. This is achieved by matching the raw mobility tracks (e.g., GPS logs) to related contextual data. The set of these enriched records

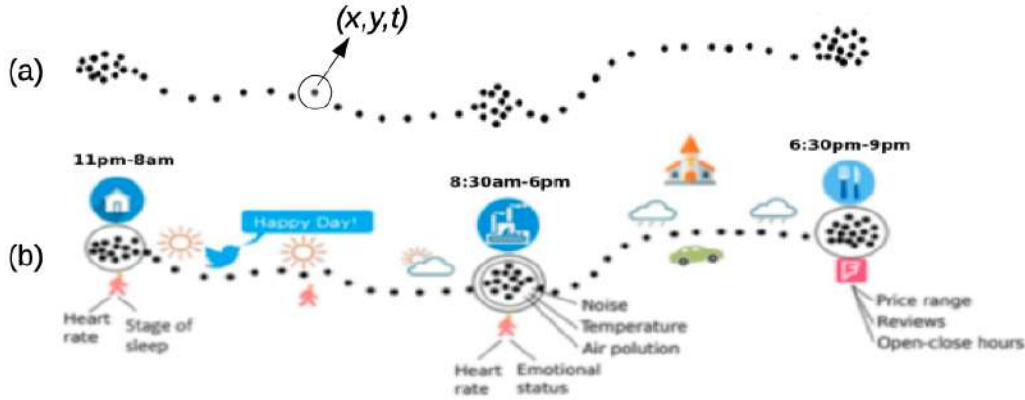


Figure 1. A raw trajectory (a) and a semantically enriched trajectory (b)

is called *semantically enriched trajectory*, or simply semantic trajectory. An example is shown in Figure 1 (b) considering the displacement of an individual from his/her home to work, and then to a restaurant. Several semantic contexts are associated with this trajectory, such as the weather condition, the way he/she travels (on foot or using a vehicle), as well as some data on some points of interest (POIs) where he/she stays.

On having such collections of semantically enriched mobility data efficiently stored in a Database (DB) system, we can query, analyze, and mine them in order to extract useful information or even knowledge for a variety of application domains. Some cases could be: (i) *Mobile/Social Marketing*: systems able to recommend, for example, cultural events or products based on users' mobility patterns; (ii) *Healthcare*: apps installed on smartphones and wearable devices that monitor users' heart, pulse, and respiratory rates alongside location data; (iii) *Natural Disasters*: social multimedia messages combined with geographic references containing spatial and temporal information unfolds the power of "citizens as sensors" to provide realtime updated data. It can be used in disaster response, timely damage assessment, as well as to efficiently carry out rescue and relief operations [Mir et al. 2018].

Some works in the literature deal with some kind of semantic trajectory [Parent et al. 2013, Alsahfi et al. 2019]. However, there is no comparative investigation regarding which persistence strategies are being proposed for this type of trajectory in order to highlight trends and best practices. This article aims to fill this gap through a survey of these strategies in order to help the development of future computational solutions for the management of semantically trajectory data. Some features are investigated and compared, such as the modeling and representation of trajectories for storage purposes, as well as DB platforms.

The rest of this paper is organized as follows. Section 2 describes another surveys related to trajectory data, Section 3 presents the applied systematic review and Section 4 presents the state-of-the-art concerning the persistence of semantic trajectories. Section 5 compares the selected studies and Section 6 is dedicated to the conclusion.

2. Related Work

The study of [Albanna et al. 2015] proposes a classification scheme for the construction and applications of semantic trajectories. This schema has 3 main classes: semantic trajectory modeling, computation and applications. For each class, approaches are mapped and challenges, research gaps and opportunities are identified.

The research by [Parent et al. 2013] addresses mobility data management and a survey of approaches to construct trajectories from motion tracks, to enrich trajectories with semantic information to enable the desired interpretations of movements; and to apply data mining to analyze semantic trajectories and extract knowledge from them. Lastly, the work examines privacy issues associated with the semantic aspects of trajectories.

The work by [Angelis et al. 2021] investigates open questions and challenges regarding work dealing with semantic trajectories and knowledge-based recommendation systems. This review emphasizes semantic technologies and machine learning techniques. The authors also propose a high-level system architecture, describing requirements for the recognition of patterns of semantic trajectories and for the inference of syntheses of visitor trajectories in cultural spaces, such as museums, suggesting new trajectories that optimize cultural experiences.

The research by [Alsahfi et al. 2019] provides a review of studies on storage, management and analysis of semantic trajectory data using data warehouse (DW) technologies. Furthermore, the authors propose a framework that provides the requirements for building the DW. Lastly, applications for this DW are treated as recommendation systems, traffic management, among others.

Different from these aforementioned surveys, this work emphasizes the persistence of semantic trajectories theme as this is a gap in the research on trajectory data. No systematic review and analysis of previous studies was found.

3. Research Methodology

In order to meet the objectives of this work, we carry out a bibliographic research methodology to obtain studies on semantic trajectory persistence. This methodology consisted of a systematic review of the state-of-the-art on the topic [Kitchenham 2004].

The systematic review process includes the definition of research questions, the search string used in the main scientific databases, and the criteria for inclusion and exclusion of works. Regarding the research questions, a main question and some secondary questions were defined, as follows:

Main Question:

RQ1 - What are the proposals for the persistence of semantic trajectories?

Secondary Questions:

RQ2 - What DB technologies are used for trajectory persistence?

RQ3 - What are the query languages or access methods used?

RQ4 - What data models are used to represent semantic trajectories?

The search strings applied to parts of the studies were the following:

Text: (“moving object“ AND database AND spatiotemporal) OR (trajectory AND database AND spatiotemporal) OR (“moving object“ AND graph) OR (multiple AND aspect AND trajectory)

Title: “moving object“ OR trajectory

Abstract: database

Finally, the following inclusion criteria were defined:

- (a) Articles written in English;
- (b) Modeling proposal includes implementation in a DB;
- (c) Works written from 2010 onwards;
- (d) Persistence of semantically enriched trajectories.

and the following exclusion criteria:

- (a) Works without implementation proposal;
- (b) Works without evaluation of results;
- (c) Works that only deal with raw trajectories.

Table 1. List of researched scientific works repositories

Repository	Found Works	Selected Works
IEEE	169	1
Wiley	55	2
ACM	144	3
Scopus	284	4
Springer	183	3
DBLP	293	1
Total	1128	14

Table 1 presents the searched repositories and the obtained results. We first found 1128 works with the application of search strings. After excluding duplicate works using the *Zotero*¹ tool, this number decreased to 391. Next, the inclusion and exclusion criteria were applied in the title and abstract, reducing the number to 146. On applying the same criteria to the full text, the final number was 14. These works are described below.

4. Semantic Trajectory Persistence Strategies

This section gives an overview of the selected works organized by persistence technology.

4.1. Object-Relational Databases

Hermes [Pelekis et al. 2014] is a pioneering DB proposal for storing raw and semantic trajectories. From the *Hermes*’ initial project, the authors have derived three independent implementations: *Hermes@Oracle*, *Hermes@Postgres* and *Hermes@Neo4J*. A set of data types based on a conceptual model that represents raw and semantic trajectories is

¹<https://www.zotero.org/>

also defined. The *trajectory* data type defines a raw trajectory through the triple (*object-id*, *trajectory-id*, *geom*), where *geom* is a three-dimensional polyline (x,y,t) that represents the object's movement. Support for semantic trajectories is achieved through the *episode* and *semantic trajectory* data types. The first represents a subtrajectory T' defined by the tuple (*defineTag*, *MBB*, *episodeTag*, *activityTag*, *T-link*) where *defineTag* defines *episode* as Stop or Move; *MBB* is a tuple (*MBR*, t_{start} , t_{end}), where *MBR* is the enveloping geometry of the spatial projection of T' and t_{start} , t_{end} the temporal projection of T' , *episodeTag* and *activityTag* contain the semantic information, and *T-link* points to the raw trajectory. At end, the *semantic trajectory* type is defined as a tuple (*o-id*, *semtraj-id*, *Tsem*), with *Tsem* storing a sequence of *episodes* ordered by time.

Another pioneering effort is the *Secondo* DB [Güting et al. 2010]: an extensible infrastructure for MODs research. It is based on the *BerkeleyDB* library for embedded DBs, comprising the following modules: (i) a *kernel* for processing SQL-like queries; (ii) an *access optimizer*; and (iii) *algebra modules*, which allow the extension of the infrastructure. One algebra in particular was developed for *symbolic trajectories* [Güting et al. 2015]. A symbolic trajectory is a sequence of pairs (t,l), where t is a time interval and l is a label describing a semantic aspect of a trajectory. Symbolic information is computed from movement or obtained from the geographic environment. Examples include names of roads and means of transport. The goal is to provide a simple model for any kind of semantic information without defining the geometric locations.

Weka-STPM [Bogorny et al. 2011] is a semantic enrichment and trajectory data mining system implemented in the *Weka* tool². Two methods are available to enrich trajectories: (i) *IB-SMoT*, which considers the presence or absence of the moving object in relevant locations; and (ii) *CB-SMoT* to semantically enrich trajectories using the maximum average velocity to identify regions of interest. With *Weka-STPM* the user is able to add semantic information to trajectories in a pre-processing step, and then explore various algorithms available in *Weka* for data mining. The solution was tested on an OR schema in the PostGIS extension of the PostgreSQL relational DB. The addition of semantics to the trajectories is carried out through the concepts of stops and movements (using *IB-SMoT* and *CB-SMoT*), whose output is persisted in two tables in the DB: STOP and MOVE. The attributes of the first table are *tid*, *stopid*, *stop_name*, *stop_gid*, *start_time*, *end_time* and *the_geometry*. The attributes of the second table are *Tid*, *Mid*, *stop_name1*, *stop_gid1*, *stop_name2*, *stop_gid2*, *start_time*, *end_time* and *the_move_geometry*.

Brandoli's work [Brandoli et al. 2022] models a DB of fishing activities in the Adriatic Sea, constructed from the trajectories of fishing vessels obtained from the AIS system³ and fish catch reports. From these data sources, trajectories are reconstructed and semantically enriched through the use of the MASTER [Mello et al. 2019] model. The solution uses the MOD *MobilityDB* [Zimanyi et al. 2020] (based on PostGIS DB), which allows representing the evolution of a value over time. In this work, the table *Trips* was modeled, which contains the columns: *MMSI* (boat identifier), *tripid* (trip identifier), and *trip*, whose type of data is *tgeompoint* (a temporal type that models the evolution of a point's position). Furthermore, the multiple semantic aspects and their attributes, such as speed, activity and volume caught, are modeled in *Trips* as columns of temporal

²<https://www.cs.waikato.ac.nz/ml/weka/>

³<https://www.marinetraffic.com/>

types, such as *ttext* and *tfloat*. The work carries out several analyzes with the objective of mapping fishing activities and inferring new knowledge useful for fisheries management, as well as to investigate the use of machine learning methods to predict the *catch per unit of effort* (CPUE) (an indicator of fisheries resource exploitation) from DB data and environmental factors, such as sea temperature and wave height.

4.2. NoSQL Graph databases

The work of [Gómez et al. 2019] proposes an OLAP graph model to favor the aggregation of trajectory data. This model is applied to the *Foursquare New York* dataset enriched with contextual information. The process of constructing semantic trajectories is carried out by replacing tuples (x,y,t) with a sequence of stops taken from the intersection of each trajectory with a collection of POIs. These stops become vertices of the graph. Each vertex has spatio-temporal attributes and is associated with one or more categories, which, in turn, are within a hierarchy that is also part of the graph. An example of a hierarchy defined in this model is *Type Category* $\rightarrow$ *Category* $\rightarrow$ *Place* $\rightarrow$ *Stop*. For persistence purposes, the Neo4j Spatial DB and the APOC library are considered.

The work of [Noureddine et al. 2021] presents a hierarchical spatial scheme that encompasses internal and external spaces, and its representation in a semantic trajectory model. The proposal is applied to real data from the *Polluscope*⁴ project, where human trajectories with multiple contextual data are collected in a crowd detection context. Such trajectories are characterized by external and geo-located data using *OpenStreetMap*⁵. Trajectories in internal spaces are generated using the simulator *VITA* [Li et al. 2016]. Both trajectories are stored in the Neo4j Spatial DB also using the APOC library. The modeling is similar to the previous proposal [Gómez et al. 2019] with the addition of the internal spaces of the POIs. On this model, queries are executed based on filters over the semantic annotations and the level of granularity of the graph vertices within the hierarchy, such as: *State* $\rightarrow$ *City* $\rightarrow$ *Road* $\rightarrow$ *Building* $\rightarrow$ *Floor* $\rightarrow$ *Room*.

The *GSM* DB emphasizes social relationships for moving objects [Zhang et al. 2016]. Persistence in GSM considers 3 graphs. The *geographical graph* models the geographic space by subdividing it into Voronoi regions [Shahabi and Sharifzadeh 2018] constructed by grouping points of trajectories closer to a POI than to any other, generating a graph where vertices represent Voronoi regions and edges represent adjacencies. The *social graph* models the relationships between moving objects. Vertices represent objects and have alphanumeric attributes. The edges represent social relationships. At end, the *motion graph* models the trajectories.

4.3. Triplestore databases

An ontology called *STEP* is proposed for semantic enrichment of trajectories treated as a series of episodes [Nogueira and Martin 2015]. This ontology holds elements for the semantic annotation of the trajectory and episodes, including quantitative, qualitative, spatial and temporal attributes. When modeling trajectories with STEP, a moving object belongs to the *Agent* class. An instance of *Trajectory*, which models semantic trajectories, can have one or more *FeatureOfInterest* (*hasFeature* property) and only one *RawTrajectory* (*hasTrajectory* property). Each *FeatureOfInterest* has one or more *Episodes* where

⁴<http://polluscope.uvsq.fr/>

⁵<https://openstreetmap.org/>

quantitative or qualitative values are noted. *FrameSTEP* [Nogueira et al. 2018] implements STEP for path persistence using *Virtuoso DB*⁶, a triplestore accessible by built-in functions or by OGC's GeoSPARQL.

The *STriDE* framework [Khalid et al. 2018] considers semantically enriched trajectories in the field of safety in civil construction. The trajectories captured through smart cameras, GPSs, smartwatches and wearable devices are semantically annotated based on a hierarchy of concepts defined through an ontology and described as RDF triples. To track changes in the shape, size and attributes of dynamic objects, *STriDE* uses the concept of *time slices*, which has 4 components: an identity, alphanumeric properties (semantic component), a time component that indicates the validity of the time slice, and a spatial component of the trajectory. On changing any component, excluding the identity, a new time slice is created inheriting the last known state of the trajectory and then these changes are applied. This newly created time slice is linked to the previous one. *STriDE* persists trajectories in the triplestore *Stardog*⁷, accessible via SPARQL and GeoSPARQL. The application of the HMM (*Hidden Markov Model*) [Liu et al. 2015] and BIM (*Building Information Modeling*) [Bulbul et al. 2014] models in the semantic trajectories allows identifying the construction sites most susceptible to unsafe movements of the users.

Torres's work presents a model of *stops-and-moves* [Alvares et al. 2007] semantically annotated [Torres et al. 2020]. The proposed trajectory representation model is based on an ontology with 3 main classes (*Trajectory*, *Stop* and *Move*), and an enrichment class (*Enrichment*). A resource t_i of the *Trajectory* class has properties *begins* and *ends*, whose values point respectively to the beginning and the end of the trajectory, belonging to the *Stop* class. In addition, t_i holds the multivalued attribute *has* whose values correspond to elements of the class *Stop* or *Move* which, in turn, can be enriched by properties of the *Enrichment* class. Persistence is via the Apache Jena triplestore⁸. This work also defines a language with search expressions over semantic trajectories that are later translated to SPARQL expressions.

An inference mechanism applied to semantic trajectories is presented in the work of Wannous [Wannous et al. 2013]. It proposes an ontology-based solution for modeling semantic trajectories by integrating data and time rules for the trajectories of seals and their activities (feeding, traveling, resting, among others) in order to identify their main feeding regions. The input data is provided by the laboratory *LIENSS*⁹, where the trajectories of these animals are captured. Ontological rules and semantically enriched trajectories are persisted in an Oracle 11g DB with the RDF Graph extension that supports hybrid queries in SQL and SPARQL. The representation of the trajectories is based on the *owlSealTrajectory* ontology in which the *Seal* class characterizes the movable object (seal). The trajectory characterized by the *Trajectory* class has one or more segments defined according to the activity of the animal in the period, modeled by the *Sequence* class, which has the properties *GeoSequence*, which stores the time intervals and the sequence of points corresponding to the condition the animal is in (on land, sea travel or diving), in addition to *Summary* (metadata) and *CTD* (*Conductivity-Temperature-Depth*).

⁶<https://www.virtuoso.com/>

⁷<https://www.stardog.com/>

⁸<https://jena.apache.org/>

⁹<https://lienss.univ-larochelle.fr/>

Table 2. Comparison of solutions for persistence of semantic trajectories

Reference	MOD	Base DB	Extension	Technology
[Pelekis et al. 2014]	Hermes	Oracle	Spatial	OR
[Pelekis et al. 2014]	Hermes	PostgreSQL	PostGIS	OR
[Güting et al. 2010]	Secondo	BerkeleyDB	Algebra module	OR
[Bogorny et al. 2011]	Weka-STPM	PostgreSQL	PostGIS	OR
[Brandoli et al. 2022]	-	MobilityDB	PostGIS	OR
[Gómez et al. 2019]	-	Neo4j	Neo4j Spatial /APOC	Graph
[Nouredine et al. 2021]	-	Neo4j	Neo4j Spatial /APOC	Graph
[Zhang et al. 2016]	GSM	Neo4j	Neo4j Spatial	Graph
[Pelekis et al. 2014]	Hermes	Neo4j	Neo4j Spatial	Graph
[Nogueira et al. 2018]	FrameSTEP	Virtuoso	-	Triplestore
[Khalid et al. 2018]	STriDE	Stardog	-	Triplestore
[Torres et al. 2020]	-	Apache Jena	-	Triplestore
[Wannous et al. 2013]	-	Oracle	RDF Graph	Triplestore
[Tamilmani et al. 2019]	-	PostgreSQL Neo4j	/PostGIS	Multi-model

4.4. Multi-model proposal

The work by [Tamilmani et al. 2019] proposes a multi-model approach, based on the PostGIS (OR) and Neo4j (Graph) DBs, which enriches the simplified geometry of raw trajectories persisted in PostGIS with the semantics lost in the simplification process, such as speed, direction, time, distance traveled, latitude and longitude. The raw trajectories are simplified according to their spatial and temporal characteristics using *Synchronized Euclidean Distance (SED)* [Meratnia and de By 2004], while a data structure named *Semantically Enriched Line simplification (SELF)* is adopted to preserve the semantics of points eliminated in the simplification process. Next, enriched simplified trajectories are modeled in the Neo4j DB. Each trajectory is represented by a graph whose vertices are classified into: (i) *ORIGIN, DESTINATION and INTERMEDIATE* and (ii) *SIDE*. Vertices have the attributes *original distance, speed, heading, time, latitude and longitude*. The SED-based simplification was applied to a set of ship trajectories.

5. Comparative

Table 2 compares the semantic trajectory persistence strategies presented in the previous section based on the research questions in Section 3. Table 3 emphasizes the data models considered for persistence and the adopted query languages. As can be seen in Table 2, there is no predominance of a particular technology, since the choice of the DB most of the time takes into account aspects of the application.

In the implementation of a MOD based on a OR DB, a trajectory is represented as an abstract data type (ADT), encapsulating spatiotemporal and semantic characteris-

Table 3. Comparison of persistence models and query language for semantic trajectories

Reference	MOD	Persistence	Query language
[Pelekis et al. 2014]	Hermes@Oracle	Relations/ADTs	SQL
[Pelekis et al. 2014]	Hermes@Postgres	Relations/ADTs	SQL
[Güting et al. 2010]	Secondo	Relations/ADTs	SQL Dialect
[Bogorny et al. 2011]	Weka-STPM	Relations	SQL
[Brandoli et al. 2022]	-	Relations/ADTs	SQL
[Gómez et al. 2019]	-	Properties graphs	Cypher
[Noureddine et al. 2021]	-	Properties graphs	Cypher
[Zhang et al. 2016]	GSM	Properties graphs	Cypher
[Pelekis et al. 2014]	Hermes@Neo4j	Properties graphs	Cypher
[Nogueira et al. 2018]	FrameSTEP	Triplestore	GeoSPARQL
[Khalid et al. 2018]	STriDE	Triplestore	SPARQL /GeoSPARQL
[Torres et al. 2020]	-	Triplestore	SPARQL
[Wannous et al. 2013]	-	Triplestore	SQL/SPARQL
[Tamilmani et al. 2019]	-	Relations/ Properties graphs	SQL/ Cypher

tics, as well as operations on its properties. Along with the proposed data structure, there must also be a manipulation language composed of set-based operations and functions in ADT, such as *Union* and *Intersect*. In addition, indexes are defined for the ADTs to increase query performance. Persistence is achieved through relationships with attributes that inherit from ADTs, and are handled through SQL. Among the analyzed MODs, Hermes@Oracle, Hermes@Postgres, Secondo and the work of [Brandoli et al. 2022] follow this strategy, while Weka-STPM implements it in part because it uses the Weka data mining tool instead of defining ADTs. PostGIS DB is adopted by most solutions.

The implementations of MODs based on triplestore present ontologically-based approaches for modeling semantic trajectories considering 3 main components: *domain ontology*, *time ontology* and *spatial ontology*, which represent basic concepts of domain contextualizing movement, space and time, as well as the relationships between them. Along with the ontologies, rules are defined. Some of them are declarative (e.g., traveling is an activity) and others are calculated from formulas on the input data. Then, a semantic integration between the ontologies is made using queries to understand temporal and spatial relationships, such as, for example, an activity happened in a certain area and during a specific time interval. The proposals emphasize the use of SPARQL and GeoSPARQL, however, they are quite heterogeneous in terms of the base DB for data persistence.

The multi-model proposal encapsulates the geometry and semantics of mobility data in different data models. The relational model keeps the raw data of GPS trajectories, which are previously cleaned to remove inaccurate data and outliers, and later persisted as a stream of spatiotemporal tuples. The graph model, on the other hand, associates several semantic properties, such as velocity, acceleration, angle of movement, density and time

interval, to parts of a raw trajectory.

Lastly, solutions based on property graphs model vertices and edges that can be annotated with properties. They are used to represent both raw and semantic trajectories. In this model, for example, the coordinates of points or episodes can be represented as vertices, and there is an edge from one point to the next one in the sequence of points of the trajectory. In addition, spatiotemporal coordinates can be included as properties, as well as other characteristics of the visited places. Hierarchical contextual data can also be defined, which allows the trajectory graph to be represented at different granularities, being very useful in statistical analysis. The representation of trajectories using graphs allows them to be stored in their native form in a DB such as Neo4j, which is a mature product with several function libraries, such as Neo4j Spatial and APOC, which is adopted by all solutions.

In short, we see that the choice for a semantic trajectory persistence approach depends on several factors, such as the application domain, the availability of the domain ontology, the level of abstraction required for the trajectory and the knowledge level of the domain designers. OR DB models based on ADTs fit a wide range of applications, allowing extensions of relational schemas and the adoption of the SQL standard, however, relational DBs are not always the best suited to handle large volumes of data, which may be the case of semantic trajectories datasets. On the other hand, ontology-based models are application domain dependent, despite being extensible, as ontologies are designed for that. The multi-model proposal supports applications that require several levels of abstraction, i.e., it performs operations along the trajectory enrichment process, considering both raw and semantic trajectories. Lastly, graph-based models allow a good representation of the path of a trajectory and its semantic properties. These last two approaches (multi-model and graph) seem to be the most promising. In the first case, we have the possibility of associating part of the representation of a semantic trajectory with the most adequate and efficient data model in terms of access. In the second case, we have the flexibility of the graph structure, which does not always require a rigid and previous schema for the data, being able to represent multiple semantic aspects.

6. Conclusion

Mobility data management and analysis had emerged in the last decade as a very active research topic, addressing issues such as clustering, integration, mining, indexing, persistence, and privacy. While previous research focused on processing raw trajectories collected from sensors, GPS devices and others, recent research focuses on methods to enrich a trajectory with more contextualized and application-oriented information. The addition of semantics brings enormous potential for the analysis of mobility-related phenomena.

This work emphasizes the persistence of semantic trajectories in order to identify related proposals. Some persistence strategies are presented in Section 5, but these proposals need to go beyond the prototyping stage to achieve high availability and allow continuous analysis of trajectories in real time, approaching the challenges of *big data* processing. In this sense, we consider it as an open topic with many possibilities for future *surveys*, such as research focused on types of indexing, summarization of long trajectories and data models for analytical purposes, among others.

References

- Albanna, B. H., Moawad, I. F., Moussa, S. M., and Sakr, M. A. (2015). Semantic trajectories: A survey from modeling to application. In Popovich, V., Claramunt, C., Schrenk, M., Korolenko, K. V., and Gensel, J., editors, *Information Fusion and Geographic Information Systems, IF&GIS 2015 - Deep Virtualization for Mobile GIS, May 18-20, Grenoble, France*, Lecture Notes in Geoinformation and Cartography, pages 59–76. Springer.
- Alsahfi, T., Alhajlah, M., and Elmasri, R. (2019). A survey on trajectory data warehouse. *Spatial Information Research*, 28.
- Alvares, L. O. et al. (2007). A model for enriching trajectories with semantic geographical information. In *15th ACM International Symposium on Geographic Information Systems, ACM-GIS*, page 22. ACM.
- Angelis, S., Kotis, K., and Spiliotopoulos, D. (2021). Semantic trajectory analytics and recommender systems in cultural spaces. *Big Data Cogn. Comput.*, 5(4):80.
- Bogorny, V. et al. (2011). Weka-stpm: A software architecture and prototype for semantic trajectory data mining and visualization. *T. GIS*, 15:227–248.
- Brandoli, B. et al. (2022). From multiple aspect trajectories to predictive analysis: a case study on fishing vessels in the northern adriatic sea. *GeoInformatica*.
- Bulbul, T., Taylor, J., and Olgun, G. (2014). A case study of embedding real-time infrastructure sensor data to bim. In *Construction Research Congress 2014*, pages 269–278.
- Gütting, R. H. et al. (2010). SECONDO: A platform for moving objects database research and for publishing and integrating research implementations. *IEEE Data Eng. Bull.*, 33(2):56–63.
- Gütting, R. H. et al. (2015). Symbolic trajectories. *ACM Trans. Spatial Algorithms Syst.*, 1(2):7:1–7:51.
- Gütting, R. H. and Schneider, M. (2005). *Moving Objects Databases*. Morgan Kaufmann.
- Gómez, L., Kuijpers, B., and Vaisman, A. (2019). Analytical queries on semantic trajectories using graph databases. *Transactions in GIS*, 23:1078–1101.
- Khalid, M. et al. (2018). Exploiting semantic trajectories using hmms and bim for worker safety in dynamic environments. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 525–530.
- Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, 33(2004):1–26.
- Li, H. et al. (2016). Vita: a versatile toolkit for generating indoor mobility data for real-world buildings. *VLDB Endowment*, 9:1453–1456.
- Liu, F. et al. (2015). Characterizing activity sequences using profile hidden markov models. *Expert Systems with Applications*, 42.

- Mello, R. et al. (2019). Master: A multiple aspect view on trajectories. *Transactions in GIS*, pages 805–822.
- Meratnia, N. and de By, R. (2004). Spatiotemporal compression techniques for moving point objects. In *Advances in Database Technology - EDBT 2004, 9th International Conference on Extending Database Technology, Heraklion, Crete, Greece, March 14-18, 2004, Proceedings*, volume 2992, pages 765–782.
- Mir, U. et al. (2018). Spatial big data and moving objects: A comprehensive survey. *IEEE Access*, 6:58835–58857.
- Nogueira, T. et al. (2018). Framestep: A framework for annotating semantic trajectories based on episodes. *Expert Systems with Applications*, 92:533–545.
- Nogueira, T. P. and Martin, H. (2015). Querying semantic trajectory episodes. In *4th ACM SIGSPATIAL International Workshop on Mobile Geographic Information Systems, MobiGIS 2015*, pages 23–30. ACM.
- Noureddine, H. et al. (2021). A hierarchical indoor and outdoor model for semantic trajectories. *Transactions in GIS*, 26:214–235.
- Parent, C., Spaccapietra, S., Renso, C., Andrienko, G. L., Andrienko, N. V., Bogorny, V., Damiani, M. L., Gkoulalas-Divanis, A., de Macêdo, J. A. F., Pelekis, N., Theodoridis, Y., and Yan, Z. (2013). Semantic trajectories modeling and analysis. *ACM Comput. Surv.*, 45(4):42:1–42:32.
- Pelekis, N. et al. (2014). *Mobility data management and exploration*. Springer.
- Shahabi, C. and Sharifzadeh, M. (2018). Voronoi diagrams. In *Encyclopedia of Database Systems, Second Edition*, pages 4580–4582. Springer.
- Tamilmani, R. et al. (2019). Modelling and analysis of semantically enriched simplified trajectories using graph databases. *Advances in Cartography and GIScience of the ICA*, 1:1–8.
- Torres, Y. et al. (2020). Stop-and-move sequence expressions over semantic trajectories. *International Journal of Geographical Information Science*, 35:1–26.
- Wannous, R. et al. (2013). *Time Integration in Semantic Trajectories Using an Ontological Modelling Approach*, volume 185, pages 187–198. Springer.
- Zhang, H. et al. (2016). Modeling and querying moving objects with social relationships. *ISPRS Int. J. Geo Inf.*, 5(7):121.
- Zimanyi, E. et al. (2020). Mobilitydb: A mobility database based on postgresql and postgis. *ACM Transactions on Database Systems*, 45:1–42.

Exploring the diversity of informal settlements in Brazil: a clustering approach to qualitative field survey data

Luis Felipe Bortolatto da Cunha¹, Flávia da Fonseca Feitosa¹, Antonio Paez²,
Carolina Moutinho Duque de Pinho¹

¹Laboratório de Estudos e Projetos Urbanos e Regionais (LEPUR) – Universidade
Federal do ABC (UFABC)
São Bernardo do Campo, SP, Brasil

² School of Earth, Environment & Society – McMaster University
Hamilton, ON, Canada

luis.cunha@ufabc.edu.br, flavia.feitosa@ufabc.edu.br,
paezha@mcmaster.com, carolina.pinho@ufabc.edu.br

Abstract. *Upgrading informal settlements and improving the living conditions of their dwellers is still a challenge. In addition, it is important to produce up-to-date information about them. This exploratory research aimed at expanding the knowledge on informal settlements through the identification, description and comparison of typologies present in Brazil. It relied on an extensive qualitative field survey applied to 157 municipalities in six regions. The territorial scope and format of data introduced methodological challenges that were addressed by using an ensemble k-modes clustering approach. The results illustrate the diversity of informal settlements through four typologies with distinctive characteristics.*

1. Introduction

UN-Habitat [2020] estimates that more than 1.6 billion people or 20% of the world's population live in inadequate housing, of which one billion reside in slums or informal settlements. Therefore, upgrading these territories and the living conditions of their dwellers remains a challenge to be faced. But in addition to the need for public policies and interventions, there is a need for accurate, consistent, and up-to-date spatial information on informal settlements that can support planning [Thomson et al. 2020, Feitosa et al. 2021, Abascal et al., 2022, dos Santos et al. 2022]. Its characteristics are diverse, varying between countries, regions, municipalities and even within the settlement. The identification of typologies helps in understanding this diversity and, consequently, in planning better strategies for mapping informal settlements.

Many studies have been developed aiming at identifying and describing typologies of deprived urban settlements in Brazil. Some of them focus on policy and/or intervention needs, such as the typologies from the National Housing Policy [Denaldi 2009], the National Policy for Regularization of Urban Land [Almeida 2018] and the Greater ABC Housing Diagnosis [CIGABC 2016]. Denaldi et al. [2018] also defined typologies of urban fabric, which were applied in precarious settlements mapping in the Metropolitan Region of Baixada Santista [Feitosa et al. 2021]. All these typologies were defined beforehand, relying on technical expertise on the subject. Other studies were based on a data-driven approach to identify the typologies, which is more similar to the

perspective embraced in this paper: Oliveira [2021] used machine learning and open data to distinguish four typologies of deprived settlements in São Paulo city, and Feitosa et al. [2022] applied a k-means clustering methodology to find 3 to 4 local typologies of informal settlements in six Brazilian regions, which showed local variations in terms of density, infrastructure precariousness, income and insertion in urban fabric.

Population Census data is often used to identify and characterize deprived urban settlements, but they have poor temporal resolution (usually 10 years) and have shown to be insufficient in capturing the diversity of informal settlements [Denaldi 2022]. Remote sensing data is also being widely used and can be updated more frequently, but they are limited to physical-territorial characteristics observable from space [Kohli et al. 2012, Kuffer et al. 2016]. Employing field survey data presents both the advantage of adopting information that is not currently present in the Census and of assimilating regional differences indirectly included from the field researcher's perspective. To the best of our knowledge, no study relied on a data-driven approach using qualitative data and the same territorial scope as the analysis here proposed.

Therefore, this research focus at exploring the diversity of informal settlements in Brazil, aiming at expanding the knowledge on its characteristics. To achieve this goal, general typologies of informal settlements are identified, described, and compared highlighting their spatial attributes. The analysis relies on an extensive qualitative field survey applied to six Brazilian regions that imposes several challenges due to its qualitative nature and magnitude. An ensemble k-modes clustering approach was adopted to identify the typologies, followed by an exploratory data analysis of the results.

2. Data

The analysis relied on a spatial dataset produced in the scope of the project 'Informal Urban Settlements' (NUI) (*Pesquisa de Núcleos Urbanos Informais no Brasil*), developed by the Institute for Applied Economic Research (Ipea) in partnership with the National Housing Secretariat (SNH) [Krause & Denaldi 2022]. The project's extent included six regions in Brazil (*Polos – Belo Horizonte, Brasília, Juazeiro do Norte, Marabá, Porto Alegre and Recife*) as an effort to represent Brazilian urban heterogeneity. The study area (Figure 1) comprised 157 municipalities, although only 137 municipalities had informal settlements reported.

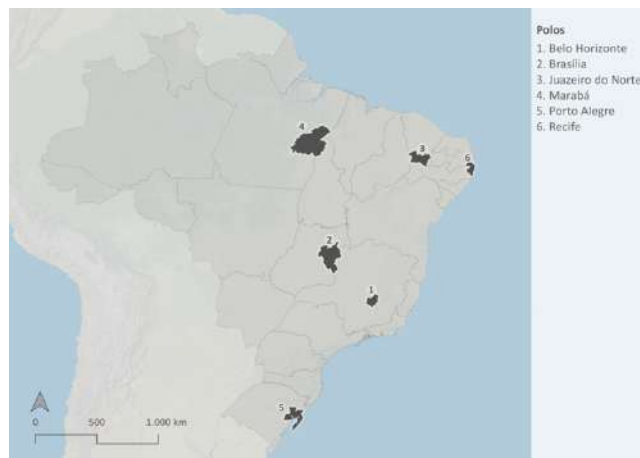


Figure 1. Study area map [Petrarolli et al. 2021]

In addition to mapping informal settlements, field researchers collected qualitative information on each of them, detailed in Table 1. The choice of variables considered the objective of the NUI project, which was to build an initial reference of settlements whose characteristics could make them eligible for federal land tenure regularization programs, and the possibility of collecting information through field survey [Petraroli et al. 2021].

Table 1. Description of the dataset

Variable	Categories	Variable	Categories
Urban contiguity: Location in relation to the urban network of the city	Central: completely inserted in the central urban network of the city	Street layout and access to parcels	Mostly present: Predominance of adequate vehicular routes that structure the occupation and most parcels have direct access to the street
	Periphery or remote: On the periphery of the urban network or completely isolated from the urban network of the city		Mostly absent: Few or no blocks are structured by vehicular routes and most of the parcels are accessed through narrow roads, alleys, or staircases
Establishment time: Approximate time of occupation	Less than 10 years		No information
	More than 10 years	Parcel subdivision: Are the parcels ordered and well-defined?	Well-defined parcels subdivision predominates
	No information		Undefined parcels subdivision predominates
Real estate dynamics: Variation in the number of dwellings in the recent period (2 years)	Increase: Rapid or slow emergence of new dwellings		No information
	Stable	Buildings distance: Is there distance between the buildings?	Present: there is a distance between the buildings (retreats and spaces between the dwellings)
	Decrease: Rapid or slow decrease in the number of dwellings		Absent: little or no distance between the buildings (retreats and spaces between the dwellings)
	No information		No information
Special Zones of Social Interest (ZEIS) intersection	Yes	Buildings quality	Adequate: popular standard dwellings predominate (consolidated and with acceptable dimensions and quality)
	No		Partial: Dwellings in different stages of consolidation and precariousness (popular standard, improvised materials, precarious constructions etc.)
	No information		Inadequate: Precarious and/or improvised dwellings predominate
Environmental Conservation Areas intersection	Yes		No information
	No	Infrastructure condition: How is the access to basic infrastructure?	Adequate infrastructure condition
Permanent Preservation Areas of water resources intersection	Yes		Partial: infrastructure is partially adequate
	No		Inadequate: Lack of basic infrastructure or very precarious existing infrastructure
Risk situation: Has there ever been an occurrence of landslide, flooding, or fire? Is it located over a contaminated area (landfill, dump etc.), gas pipeline, water mains, or under a high voltage line?	Yes		No information
	No		
	No information		
Risk susceptibility: Is it, or part of it, inside an area of medium or high susceptibility to risk?	Yes		
	No		
	No information		

3. Methodology

Identifying typologies, from a data mining perspective, consists in assigning labels to observations in such a way that those with the same label are similar to each other and those with different labels are dissimilar in some sense. Since the labels are not available beforehand, this process is known as unsupervised classification or clustering – a well-established statistical method for finding groups in data, used in many contexts and in many disciplines. It is essentially an exploratory analysis, which can uncover hidden patterns in data [Jain et al. 1999, Bandyopadhyay & Saha 2013, Hennig et al. 2015].

The k-means algorithm is arguably the most popular clustering method because it is computationally easy, fast, and memory-efficient. But on its original formulation, it employs the Euclidean distance as a measure of dissimilarity and minimizes the cost function by changing the means of clusters, being only applicable to numeric values [MacQueen 1967]. Based on the k-means, Huang [1998] proposed the k-modes algorithm, designed to derive clusters from categorical data by using the simple matching distance as dissimilarity measure and minimizing the cost function by changing the modes of clusters. The simple matching distance between two vectors can be defined by the total mismatches of the corresponding attribute categories of the two objects.

As any machine learning algorithm, it has several limitations. The most relevant one is that it may not converge to an optimal solution because it is dependent of the choice of the initial cluster (seed). For this reason, the k-modes algorithm was placed into a cluster ensemble, which is a methodology for combining different clustering results into a single consensus cluster. Cluster ensembles are highly recommended as they have shown to be better than any standard clustering algorithm at improving the accuracy and robustness of clustering and they can be applied to the output of the k-modes algorithm – in fact, the cluster ensemble can be treated as a problem of clustering categorical data (using the output of several clusters as features) [Strehl & Ghosh 2002, Ayad & Kamel 2010, Ghosh & Acharya 2011, Boongoen & Iam-On 2018].

Figure 2 shows the methodology steps, which included: (1) Exploratory data analysis to choose the number of clusters, (2) Clustering ensemble model (k-modes and consensus clustering), and (3) Descriptive and comparative analysis of the output.

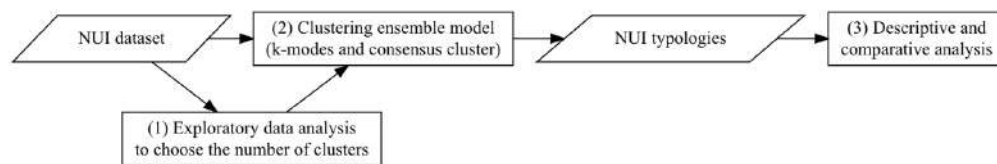


Figure 2. Methodology steps

Since the number of clusters (k) in the k-modes algorithm must be defined beforehand, an attempt of addressing this problem was made by creating an elbow plot showing the dissimilarity decay by k (measured as the mean and standard deviation simple matching distance across 100 models with different seeds). Using this plot as reference, some alternatives were tested and analyzed subjectively for the definition of the final number of clusters. Afterwards, the k-modes algorithm [Weihs et al. 2005] was

applied 1000 times with different seeds to a random subsample of 80% of the NUI dataset. The results were introduced into a consensus clustering function [Chiu & Talhouk 2018], that relabels the clusters to make them comparable and find the majority voting across each observation, resulting in the final NUI typologies. The NUI typologies were then analyzed, described, and compared. All computation was done with R statistical computing software and the code is available on GitHub¹.

4. Results and discussion

A first assessment of the dataset aimed at choosing the number of typologies (or clusters). Figure 3 shows the dissimilarity decay as the number of clusters increase, indicating that the clustering methodology can reduce the within-cluster dissimilarity. Nonetheless, it does not highlight one unique solution. Therefore, a subjective analysis of the output considering 3 to 6 clusters was made, leading to the choice of $k = 4$ for its higher coherence and interpretability.

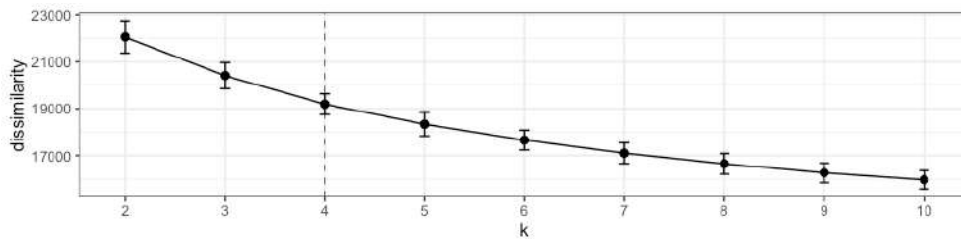


Figure 3. Elbow plot

Table 2 shows the proportion of typologies by region, Table 3 presents a summary of the characteristics of each typology, Figure 4 illustrates a settlement of each typology, including an aerial view and a street view, and Figure 5 shows maps with the spatial distribution of the typologies by region. The analysis of this information enabled a preliminary description of each typology:

- **Typology 1 (T1)** includes informal settlements that usually have appropriate street layout (95%), well-defined parcel subdivision (89%), distance between buildings (86%), and adequate buildings quality (73%). They are usually located at peripheric locations (71%), and few had risk situation reported (19%) or risk susceptibility mapped (12%), even though only 18% have adequate infrastructure. This is the most common typology (40% of all settlements mapped), and it is predominant in Brasília (93%), Juazeiro do Norte (75%), Marabá (52%), and Porto Alegre (48%), and significantly present in Belo Horizonte (30%) and Recife (16%) as well. Figure 4a shows a settlement at the periphery of the urban network (aerial view and street view), which has a regular street layout but lacks basic infrastructure (paved roads) in part of it. Figure 5 illustrates that it is widespread in Brasília and Juazeiro do Norte, and present mostly as isolated settlements in other regions.

¹ <https://github.com/luisfelipebr/geoinfo2022>

- **Typology 2 (T2)** informal settlements are characterized by their central location (90%), lack of distance between buildings (85%), being often mapped as special zone of social interest in local plans (76%), and presence of risk susceptibility (59%). The condition of buildings (10% inadequate) and infrastructure (9% inadequate) are comparatively better than T3 and T4. This typology is more common in Belo Horizonte (35%), Porto Alegre (34%), and Recife (43%) regions. Figure 4b displays a very high-density settlement, in which several buildings have more than one floor built, and the terrain is steep, which can lead to risk situation and make access to the parcels difficult (made by ramps and stairs). Figure 5 highlights the location of these settlements, usually at the central municipality.
- **Typology 3 (T3)** main characteristics are the presence of permanent preservation areas of water resources (97%) and incidence of risk situation (86%), which are often associated. They are mostly located in central areas (74%) and more than half of them had an increase in the number of dwellings in the recent period (58%). This typology stands out in Belo Horizonte (31%) and Marabá (23%) regions. Figure 4c demonstrates a settlement close to water resources, which intersects with environmental conservation areas and permanent preservation areas of water resources, and with risk situation and lack of basic infrastructure reported. This typology is not as usual as T1 and T2 (15% of all informal settlements mapped) and is mostly visible in Belo Horizonte region (Figure 5).
- **Typology 4 (T4)** settlements have the worst basic infrastructure condition (68% are inadequate) and quality of buildings (only 3% are adequate). Its parcel subdivision is not well-defined (15%) and there is a lack of information about risk situation (63%) and risk susceptibility (46%), which might be related to the fact that a lot of these settlements were recently established (29%) and the number of dwellings in recent period increased in more than half of the settlements in this typology (66%). This typology appears mostly in Recife region, where 33% of the settlements are of Typology 4, and they are usually not located at the central municipality (Figure 5). Figure 4d shows an isolated occupation in Marabá region, with inadequate buildings quality and infrastructure condition.

Table 2. Proportion of typologies by region

Region	T1 (n = 1995)	T2 (n = 1567)	T3 (n = 733)	T4 (n = 673)
Belo Horizonte	0,30	0,35	0,31	0,04
Brasília	0,93	0,03	0,03	0,01
Juazeiro do Norte	0,75	0,05	0,13	0,07
Marabá	0,52	0,10	0,23	0,15
Porto Alegre	0,48	0,34	0,07	0,11
Recife	0,16	0,43	0,08	0,33

Table 3. Proportion of settlements in each category by variable and typology

Variable	Categories	All settlements (n = 4968)	T1 (n = 1995)	T2 (n = 1567)	T3 (n = 733)	T4 (n = 673)
Urban contiguity	Central	0,56	0,29	0,90	0,74	0,41
	Periphery or remote	0,44	0,71	0,10	0,26	0,59
Establishment time	Less than 10 years	0,13	0,14	0,06	0,10	0,29
	More than 10 years	0,86	0,85	0,94	0,89	0,69
	N.A.	0,01	0,01	0,00	0,01	0,02
Real estate dynamics	Increase	0,33	0,32	0,10	0,58	0,66
	Stable	0,66	0,68	0,89	0,40	0,33
	Decrease	0,01	0,01	0,02	0,02	0,01
	N.A.	0,00	0,00	0,00	0,00	0,00
Special Zone of Social Interest	Yes	0,42	0,24	0,76	0,40	0,16
	No	0,42	0,52	0,17	0,50	0,63
	N.A.	0,16	0,24	0,07	0,10	0,21
Environmental Conservation Areas	Yes	0,22	0,31	0,10	0,24	0,19
	No	0,78	0,69	0,90	0,76	0,81
Permanent Preservation Areas	Yes	0,49	0,37	0,41	0,97	0,47
	No	0,51	0,63	0,59	0,03	0,53
Risk situation	Yes	0,33	0,19	0,34	0,86	0,16
	No	0,43	0,58	0,53	0,04	0,21
	N.A.	0,24	0,23	0,13	0,10	0,63
Risk susceptibility	Yes	0,34	0,12	0,59	0,35	0,37
	No	0,18	0,11	0,28	0,16	0,18
	N.A.	0,49	0,77	0,14	0,48	0,46
Street layout	Mostly present	0,76	0,95	0,59	0,89	0,46
	Mostly absent	0,24	0,05	0,41	0,10	0,54
	N.A.	0,00	0,00	0,00	0,00	0,00
Parcel subdivision	Well-defined	0,66	0,89	0,59	0,67	0,15
	Undefined	0,33	0,11	0,41	0,32	0,85
	N.A.	0,00	0,01	0,00	0,00	0,00
Buildings distance	Present	0,57	0,86	0,14	0,60	0,63
	Absent	0,43	0,13	0,85	0,40	0,36
	N.A.	0,00	0,01	0,00	0,00	0,00
Buildings quality	Adequate	0,47	0,73	0,41	0,28	0,03
	Partial	0,32	0,13	0,42	0,40	0,60
	Inadequate	0,12	0,05	0,10	0,21	0,25
	N.A.	0,09	0,08	0,08	0,11	0,13
Infrastructure condition	Adequate	0,20	0,18	0,32	0,11	0,03
	Partial	0,55	0,68	0,52	0,65	0,16
	Inadequate	0,18	0,09	0,09	0,18	0,68
	N.A.	0,07	0,06	0,07	0,06	0,13



Figure 4. Typologies examples (Google Satellite and Google Street View)

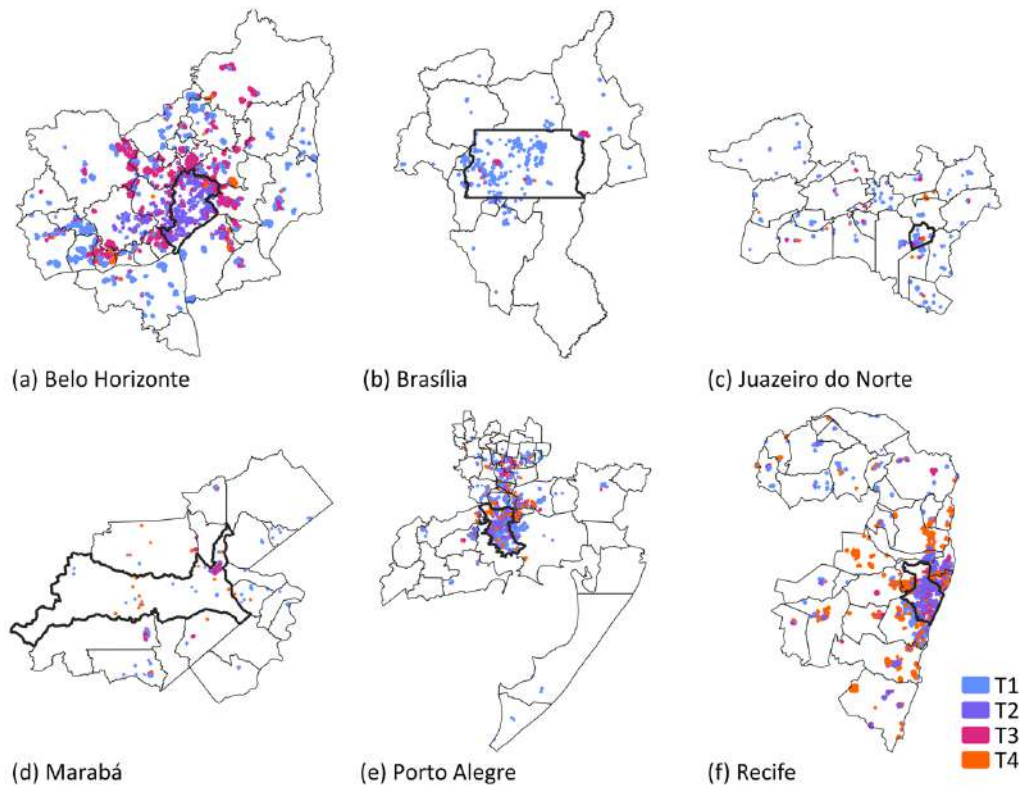


Figure 5. Spatial distribution of typologies by region

The typologies identified through the clustering approach are quite distinct and display four different facets of the precariousness present in informal settlements in Brazil. The analysis benefited from the subjective perception of the mapping experts since the definition of adequate/inadequate varies according to the context of each region. The results show how location is a relevant characteristic of these settlements, expressed through the urban contiguity variable, and highlight regional differences, with some regions being less diverse and others more diverse.

5. Conclusion

The exploratory analysis presented helped to emphasize the diversity of informal settlements in Brazil. The results show four typologies with distinctive characteristics, which were further described and compared. These typologies are widespread across the six regions, but in varying proportions. By mapping the typologies, it was also possible to understand the geographical nature of precariousness in each region.

The results can feed into the mapping of informal settlements and the process of planning public policies and interventions. Further research may explore other methods for choosing the number of clusters, arriving at a consensus cluster, or deriving the clusters altogether. Acknowledging the limitations of the clustering technique adopted and of qualitative data, the analysis could be refined by adopting mixed data from multiple data sources. The methods and results are openly available, which may lead to additional insights on informal settlements diversity.

Acknowledgements

Ipea and SNH for providing the dataset used in the analysis. The first author was partially funded with a scholarship provided with the support of the Government of Canada.

References

- Abascal, A., Rothwell, N., Shonowo, A., Thomson, D., Elias, P., Elsey, H., ... & Kuffer, M. (2022). “Domains of deprivation framework” for mapping slums, informal settlements, and other deprived areas in LMICs to improve urban planning and policy: A scoping review. *Computers, environment and urban systems*, 93, 101770.
- Almeida, G. M. J. A. de (2018). Estudo da demanda para regularização fundiária de assentamentos urbanos no âmbito do Programa Papel Passado. Santo André: Ed. UFABC, 2018.
- Ayad, H. G., & Kamel, M. S. (2010). On voting-based consensus of cluster ensembles. *Pattern Recognition*, 43(5), 1943-1953.
- Bandyopadhyay, S., & Saha, S. (2013). Unsupervised classification: similarity measures, classical and metaheuristic approaches, and applications. Berlin: Springer.
- Boongoen, T., & Iam-On, N. (2018). Cluster ensembles: A survey of approaches with recent extensions and applications. *Computer Science Review*, 28, 1-25.
- Chiu, D. S., & Talhouk, A. (2018). diceR: an R package for class discovery using an ensemble driven approach. *BMC bioinformatics*, 19(1), 1-4.
- CIGABC – Consórcio Intermunicipal do Grande ABC (2016). Diagnóstico Habitacional Regional do ABC. (Report).
- Denaldi, R. (2009). Assentamentos precários: identificação, caracterização e tipologias de intervenção. In Denaldi, R. & Rosa, J. (Eds.). *Curso à Distância: Planos Locais de Habitação de Interesse Social*. Brasília, Ministério das Cidades/Aliança de Cidades.
- Denaldi, R., Petrarolli, J., Gonçalves, G., Moraes, G. (2018). Tecidos Urbanos e a Identificação de Assentamentos Precários na Região Metropolitana da Baixada Santista. *Proceedings of III URB Favelas*, Salvador, BA, Brasil.
- Denaldi, R. (2022). Identificação e caracterização da precariedade habitacional: desafios conceituais e metodológicos. In Krause, C., Denaldi, R. (Eds.) (2022). *Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional*. Ipea [Book in press].
- dos Santos, B. D., de Pinho, C. M. D., Oliveira, G. E. T., Korting, T. S., Escada, M. I. S., & Amaral, S. (2022). Identifying Precarious Settlements and Urban Fabric Typologies Based on GEOBIA and Data Mining in Brazilian Amazon Cities. *Remote Sensing*, 14(3), 704.
- Feitosa, F., Vasconcelos, V., Pinho, C., Silva, G., Gonçalves, G., Danna, L., & Lisboa, F. (2021). IMMerSe: An integrated methodology for mapping and classifying precarious settlements. *Applied geography*, 133, 102494.

- Feitosa, F., Gonçalves, G., Cunha, L. (2022). Aglomerados Subnormais e Núcleos Urbanos Informais: Uma Análise Comparativa. In Krause, C., Denaldi, R. (Eds.) (2022). Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional. Ipea [Book in press].
- Ghosh, J., & Acharya, A. (2011). Cluster ensembles. *Wiley interdisciplinary reviews: Data mining and knowledge discovery*, 1(4), 305-315.
- Hennig, C., Meila, M., Murtagh, F., & Rocci, R. (Eds.). (2015). *Handbook of cluster analysis*. CRC Press.
- Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, 2(3), 283-304.
- Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3), 264-323.
- Krause, C., Denaldi, R. (Eds.) (2022). Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional. Ipea [Book in press].
- Kohli, D., Sliuzas, R., Kerle, N., & Stein, A. (2012). An ontology of slums for image-based classification. *Computers, Environment and Urban Systems*, 36(2), 154-163.
- Kuffer, M., Pfeffer, K., & Sliuzas, R. (2016). Slums from space—15 years of slum mapping using remote sensing. *Remote Sensing*, 8(6), 455.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability* (pp. 281-297).
- Oliveira, L. (2021). The diversity of deprived areas: applications of unsupervised machine learning and open geodata (Master's thesis, University of Twente).
- Petrarolli, J. et al. (2021). Produto 11 – Relatório de consolidação das pesquisas de campo sobre aspectos físicos e territoriais de núcleos urbanos informais. Pesquisa de Núcleos Urbanos Informais no Brasil (Report).
- Strehl, A., & Ghosh, J. (2002). Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of machine learning research*, 3(Dec), 583-617.
- Thomson, D. R., Kuffer, M., Boo, G., Hati, B., Grippa, T., Elsey, H., ... & Kabaria, C. (2020). Need for an integrated deprived area “slum” mapping system (IDEAMAPS) in low-and middle-income countries (LMICs). *Social Sciences*, 9(5), 80.
- UN-Habitat (2020). *World Cities Report 2020: The Value of Sustainable Urbanization*. UN-Habitat, Nairobi.
- Weihs, C., Ligges, U., Luebke, K. and Raabe, N. (2005). *klaR Analyzing German Business Cycles*. In Baier, D., Decker, R. and Schmidt-Thieme, L. (eds.). *Data Analysis and Decision Support*, 335-343, Springer-Verlag, Berlin.

Application of GEOBIA for urban area classification using optical images from the PE3D project

Sabrina O. Gomes¹, Ana L. Candeias²

^{1,2}Department of Cartographic Engineering – Federal University of Pernambuco (UFPE)

`o.sabrina.gomes@gmail.com, analucia@ufpe.br`

Abstract. *The urban space is complex and has great variability of targets. The methods of extracting information, automated or manual, present limitations and difficulties. Although high resolution images allow better visibility of urban targets, there is a difficulty in automatically discriminating targets. The use of auxiliary layers and pre-processing in the image can help in the classification process. This work presents a semi-automation of feature extraction to assist in the vectorization of blocks, streets and lots. A subdivision in Caruaru, Pernambuco, was used as a test area. The results can help in the Processed Product block and saved on the basis of COMPESA (Companhia Pernambucana de Saneamento).*

1. Introduction

COMPESA (Companhia Pernambucana de Saneamento) needs to update its registration for the purposes of installing and charging its services. It provides services to the entire state of Pernambuco. It is divided into two main directorates: the DRM, which takes care of the Metropolitan Region of Recife, and the DRI, which works with the interior region. This work is focused on the semi-automation of feature extraction to assist in the vectorization of blocks, streets and lots. The results will compose the Processed Product block and saved on the basis of COMPESA.

Remote sensing and image processing will provide the resources for this automation. This is possible because advances in orbital sensors related to improving spatial, spectral, radiometric and temporal resolution have enabled improvements in urban studies, such as better identification of targets at large scales. It is important to note that linked to this, there is a cost to obtain these images.

The Pernambuco Tridimensional (PE3D) program is an initiative motivated after the floods that hit the Mata Sul regions of the State of Pernambuco in 2010 and 2011. Due to this, mapping of the affected areas was carried out in order to control future floods, this mapping was possible from laser scans, where the Secretary of Water and Energy Resources was responsible for the project. PE3D is one of the goals that were included in the Pernambuco Water Sustainability Program which aims to expand access to drinking water and sanitation for the population. In this way, PE3D comes with the objective of generating the entire mapping of the State, making these products available for free.

A spatial database obtained from aerophotogrammetric coating and laser profiling brings a positive development to the region, generating a solid base for the academic and business field. One of the products that were generated from the program were the Orthophotos at a scale of 1:5000, which were made from aerophotogrammetric coverage, with a spatial resolution of 50cm, which allows a good distinction of targets in intra-urban regions. Therefore, these products bring the possibility of studies in urban scenarios from image information extraction techniques.

According to Meneses et al. (2012), image processing consists of performing mathematical operations on the data, aiming at their transformation into images with better spectral and spatial qualities and that are more appropriate for a given application. The Geographic Object-Based Image Analysis (GEOBIA) is a technique that aims to individualize the object from clusters of pixels that are similar in terms of spectral and spatial aspects, as well as considering attributes in terms of shape, size and others (Blaschke et al., 2014).

To perform the individualization of objects, it is important to use some techniques, such as data mining, which allows patterns to be discovered for data analysis from processing through algorithms. According to Wintten et al. (2011), data mining is an automatic or semi-automatic process of discovering patterns in data. Usually the data are grouped in considerable amounts, as in GEOBIA, so the discovery of the patterns must be significant in order to benefit the image classification.

Discovering patterns in data allows you to relate the data to a certain class. For this, it is possible to determine a set of rules for classification through a decision tree.

This classification structure is composed of leaf nodes that contain the classes and the intermediate nodes are the test-based attributes having a branch for each possible output.

Thus, this work has the general objective of determining a methodology for extracting urban targets and generating thematic mapping of the urban environment to aid decision-making in the territory.

2. Study Area

The test area is in Caruaru, Pernambuco and was delimited considering some aspects of the region. The analyzed point was the diversity of targets in the joint so that they were able to be distinguished by the human eye (Figure 1). The importance of applying the test area in an urban environment is that there is significant heterogeneity in land use and occupation. Therefore, the targets were considered: vegetation, roofs, exposed soil and road.

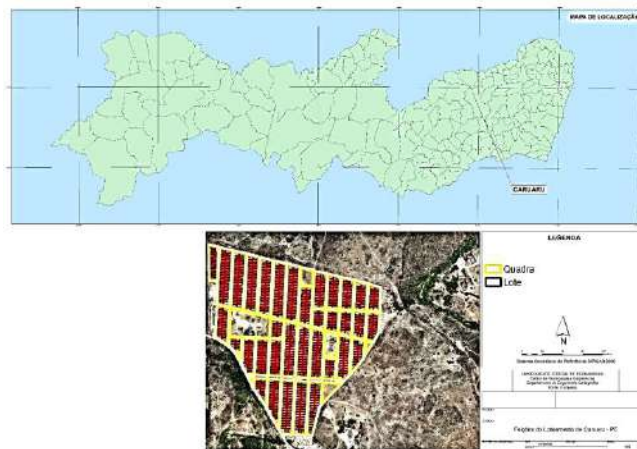


Figure 1. Study area map.

The area is characterized by an urban environment, called subdivision area, which has the characteristic of having an orderly growth in terms of its urbanization, with the presence of residential properties where most are characterized by houses, contains paved patios, a portion of green area and exposed soil area. In addition, the areas around the subdivision will also be considered in the study, as they show other characteristics such as denser vegetation.

3. Material

3.1. Input Data

- Matrix (raster):

The acquired images were made available by Companhia Pernambucana de Saneamento (COMPESA) from the PE3D.

Orthophoto of the Municipality of Caruaru in GeoTIFF format with a scale of 1:5000, spatial resolution of 50cm, radiometric resolution of 8bits, referenced in the Projected Coordinate System UTM SIRGAS 2000 24S time zone.

- Vectors:

The territorial limits of the State of Pernambuco and the Municipality of Caruaru were used, in shapefile format (.shp) acquired from the vector base made available free of charge by the Brazilian Institute of Geography and Statistics (IBGE).

The vectorization of the image features was acquired from Companhia Pernambucana de Saneamento (COMPESA), made available by the Registration and Geoprocessing Management (GCG) which is responsible for making the definitions of the image features from vectors called Blocks, Lots and Streets.

3.1. Software and Applications

TerraView is an application developed from an open geoprocessing software library called TerraLib, built by the National Institute for Space Research - INPE, which aims to support the development of customized geographic applications. TerraView is a geographic data viewer that allows an advanced analysis of vector and matrix data, where this data can be stored in a Database Management System. TerraView has the following main objectives:

- Be an easy-to-use application, with data visualization and analysis capabilities.
- Demonstrate the use of the TerraLib library.

It is important to point out that the TerraView application only manipulates existing data, that is, it does not allow constructions of geographic data to be made. One of TerraView's functions is to manage the database and manipulate the images, while the GeoDMA tool performs the entire process necessary to classify the image.

The Geographical Data Mining Analyst known as GeoDMA is a plugin developed for the TerraView system and works to perform image analysis from data mining techniques, being a powerful tool for applications of remote sensing techniques. The plugin has the following main activities:

- Segmentation.
- Attribute extraction.
- Visualization of attributes.
- Classification.
- Visualization.

Spring is a software developed by INPE (National Institute for Space Research) and DPI (Image Processing Division) with several partners, which has the functionality

of a GIS (Geographic Information System) with image processing, analysis and spatial, terrain modeling and others.

The main objectives of the software are to enable the following activities:

- Construction of a rapid learning Geographic Information System, accessible and possible to apply in different areas of study.
- Unification of a Geoprocessing and Remote Sensing environment.

4. Methods

4.1. Class Definition

For the definition of the classes, a previous visual analysis of the characteristics of the area was carried out.

The subdivision area has four visibly striking classes, the built-up areas characterized by buildings and roads and the unbuilt areas, characterized by vegetation and exposed soil.

The blocks are defined from the grouping of lots, defined by the roofs and next to this. There are the segments of the roads, which make it possible to improve the definition of the blocks. These classes were defined to better characterize the study environment.

4.2. HSI Transformation

To perform the HSI transformation, the three bands of the visible range of the acquired image were used, namely R (red), G (green) and B (blue), the expected result is that each color pixel of the image will have a corresponding to the HSI space.

The histogram indicates the image quality in terms of contrasts and brightness. It is possible to notice that in (Figure 2), the histogram of I (intensity) of study area (Figure 1), and Figure 3 and Figure 4 show respectively, H (Hue – Hue) and S (Saturation).



Figure 2. Histogram I



Figure 3. Histogram H



Figure 4. Histogram S

4.2. Segmentation

To visualize and handle the image in the TerraView software, it is necessary to create a database and thus import the image into the software. Through the GeoDMA tool that performs the classification process.

The first step in applying GEOBIA is to segment the image. According to Paulo et. al (2012), segmentation produces vector representations, which are generated from pixels with similar responses. That is, the image will be divided by regions. These regions are sets of pixels that appear uniformly in the image. The same author states that the quality of this grouping depends on the technique used. The parameters that are defined by the user and the characteristics of the bands available in the image.

To perform the segmentation, it is mandatory to insert parameters that control the process, the Euclidean distance and minimum area. The segmentation was performed using the growth method by regions. The evaluation of similarity between pixels is done through the Euclidean distance. It is the analysis parameter of the growth method by regions available in GeoDMA. This parameter defines whether two pixels will be aggregated in a region from the distance that was defined, a low value means that the algorithm will create regions with little variability, while if the values are high, the regions will have more variability in terms of grey level.

The parameter called minimum area defines what will be the smallest area that will be created, if any object is initially smaller than the one defined by the user, and then it will be joined with some neighbor. Both aspects must be tested from combinations until the best image positioning is found. The values used for the segmentation parameters were: *Minimum Area=10* and *Euclidean Distance=20*

4.2. Classification by Decision Tree

Attribute extraction allows various attributes of objects to be obtained, such as spectral and spatial characteristics. The extraction considers the image and objects as input data, thus storing them in the TerraView database for future use. Furthermore, when we extract these attributes, we can visualize them from scatter plots, where we can correlate the attributes and thus visualize the separability of the data.

In GeoDMA, spatial attributes are derived from polygons, as well as spectral attributes, which can be combined with spatial information, having the nomenclature *rp*. In addition to these, there are still attributes that can be extracted from cells, as well as attributes for multitemporal analysis. For the work, it will not be necessary to use the cell-type and multi-temporal attributes.

Once the attribute extraction step is performed, samples are selected for the classification process. To perform the supervised classification, it is necessary for the user to select samples that represent certain classes present in the image.

Training is the process in which samples are selected and assigned to a certain user-defined class, that is, objects are selected and associated with the created classes. Four types of classes were created: Roads, Exposed Soil, Vegetation and Roofs. The attributes considered to perform the extraction were related to color, shape, texture and other visual attributes.

After the training stage, we proceed to the classification stage, in which we include all the attributes that were extracted and the samples for classification. One of the necessary parameters that is requested by the software is for the classification through the decision tree called “minimum number of objects per sheet”. This parameter indicates the limit of objects (segments) that will be in the sheet so that it is understood as valid during the training stage, this parameter does not directly prevent problems in the classification.

From this, the software returns the decision tree, created from all the processing done by the user, in this way the user validates the model, and can understand the thresholds used for classification. Körting et. al. (2009), states that the use of GeoDMA is an advantage for the process of extraction, normalization and storage of data attributes due to its ease, and that later can be used in other analyses.

5. Description and discussion of results

5.1. Stacking Evaluation

The HSI image allowed to highlight the targets of interest, roads, roofs and vegetation, when stacked with the indices, it was possible to better highlight each information. The objective of performing the stacking was to unite the information from the HSI image with the GRVI and VARI indices and thus provide an improvement in the visualization of the targets.

Below you can see (Figure 5), it shows only the HSI image viewed in the RGB channels. It is associated to R channel, H to G, and S to B. Image highlights targets, such as denser vegetation, characterized by the green color. The exposed soil and some roof targets have similar colors, which would make the classification process difficult to separate them. The road targets present the color in a different tone from the others, which would help not to generate doubts to classify them and collect the samples.

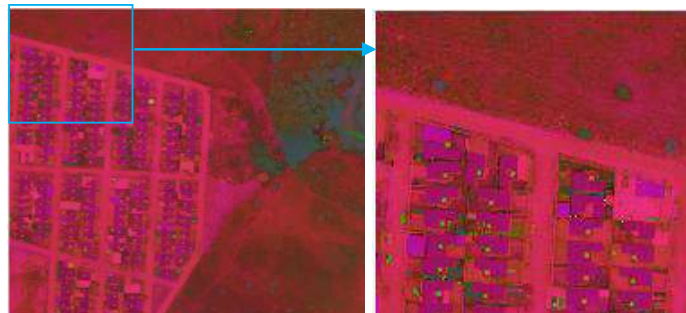


Figure 5. HSI image before stacking with Indexes

With the use of vegetation indices, it was possible to highlight targets of interest for the classification, especially roofs and vegetation, we can see the highlight of denser vegetation and roofs in the GRVI index and a little emphasis on the roofs in the VARI index.

According to Silva (2020), the VARI and GRVI indices are not potential for the identification of intra-urban vegetation. However, the GRVI presents a better identification of the targets as being building coverage, highlighting for objects of the

ceramic roof type, while the VARI for the ceramic roof type did not classify roofs or parts of roofs with darker shades.

However, when stacking the HSI image with the vegetation indices, an improvement in the visualization and highlighting of the targets is noticeable. It is possible to notice that the exposed soil was uniformly characterized throughout the area, the densest vegetation located in the right region of the image remained highlighted, and the targets of roofs and public places were notable highlights.



Figure 6. HSI Image with indexes

The same occurs with the original image, the same when stacked with the indices presents an improvement in the highlights of the targets, mainly in the vegetation. It is possible to notice that some targets present similarity in their gray level, public spaces and exposed soil, which could make the classification process difficult.



Figure 7. Original Image with Indexes

5.2. Rating Assessment

It is important to assess the differences between the classification of the original image and the HSI image, in order to observe the differences between the results and if there was an improvement to the classification process of the targets and their classification.

The amount of samples collected for the original image were the Road class that had 118 samples collected, where 70 will be used for training and 48 for tests, the same happens for the other classes, Exposed Soil with 137 samples, Roofs with 514 and Vegetation with 142 samples. These values are distributed by the software itself, seeking a proportion of 60% for training and 40% for testing the classification tree.

The number of samples for the HSI Image were, Roads 108 samples, Exposed Soil 138, Roofs 497 and Vegetation 129 samples. The objective was to use the maximum possible amount of equal samples between the two images, but as mentioned earlier, the original image had more segmented targets, which facilitates the sample collection process, while the HSI image did not have the same result of super-segmentation.

As already mentioned, the segmentation step can be one of the main causes of assigning a target to a class to which it does not belong, since segmentation can group different targets to the same segment. One way to assess the quality of the results obtained is through the confusion matrix, Tables 1 and 2. The matrix correlates the information of the true data, that is, a set of samples that were used in the training, with the final data that are classified. From the confusion matrix, it was possible to analyze the best and worst classified classes, the errors of omission (EO) which is in the line of the confusion matrix and measures the accuracy of the classification and the commission errors (EC) located in if in the column of the confusion matrix that measures the reliability of a class matches the truth.

It is possible to analyze that from Table 1 the Roofs and Vegetation classes obtained a good classification, observing the diagonal of the matrix, the Roofs class had 196 samples truly belonging to the class out of a total of 204, that is, only 8 samples were not assigned to Is it over there. The same occurs in the Vegetation class, of the 57 samples and 50 were truly belonging to the class, while 7 samples were not assigned to it.

To analyze the accuracy of the classification, the calculation for omission errors was used, which showed the values of Road 35.89%, Exposed Soil of 30.76%, Roofs 4.85% and Vegetation 9.09%. While to calculate the reliability of a class being correctly classified, it was from the calculation of commission errors, each class corresponded to the following value, Road 47.92%, Exposed Soil 19.64%, Roofs 3.92% and Vegetation 12.28%.

The global accuracy can be calculated as follows, the sum of the main diagonal elements, which indicates the total of correctly classified samples, in this case 291, divided by the total number of validations, 365. The global accuracy will result in 79.72%. According to Meneses et. al. (2012), the overall accuracy can vary on a scale from 0 to 100%.

Table 1. Original Image Confusion Matrix

-	Road	Exposed Soil	Roofs	Vegetation	Total	EO %	EC %
Road	25	14	8	1	48	35,89	47,92
Exposed Soil	6	45	1	4	56	30,76	19,64
Roofs	7	1	196	0	204	4,85	3,92
Vegetation	1	5	1	50	57	9,09	12,28
Total	39	65	206	55	365		-

The results of the HSI image can be observed in Table 2, it is possible to analyze that the Roofs and Vegetation classes obtained a good classification, observing the diagonal of the matrix, the Roofs class had 198 samples collected, where 176 truly belong to the class and only 22 samples were not assigned, Vegetation had 36 samples

truly belonging to the class out of a total of 52, that is, only 16 samples were not assigned to it.

To analyze the accuracy of the classification, the calculation for errors of omission was used, which showed the values of Road 64.00%, Exposed Soil of 34.21%, Roofs 19.63% and Vegetation 48.57%. While to calculate the reliability of a class being truly classified, it was from the calculation of commission errors, each class corresponded to the following value, Road 79.07%, Exposed Soil 57.62%, Roofs 11.11% and Vegetation 30.77%.

The global accuracy was calculated by the sum of the main diagonal elements (246) divided by the total validation number (352), resulting in an overall accuracy of 69.89%.

Table 2. Contrast HSI Image Confusion Matrix

-	Roofs	Exposed Soil	Roofs	Vegetation	Total	EO %	EC %
Road	9	2	29	3	43	64,00	79,07
Exposed Soil	0	25	8	26	59	34,21	57,62
Roofs	15	2	176	5	198	19,63	11,11
Vegetation	1	9	6	36	52	48,57	30,77
Total	25	38	219	70	352	-	-

It is possible to observe from the images below that the Roof classes in both images were visually well defined in the classification, the same occurs for the Roads class. The segment of the same was well defined, but it is possible to notice that there are some "holes" in the classification, where sections of the Road class were assigned to other classes, such as exposed soil and vegetation, it is possible to notice that this occurs more frequently in both regions. images.

The attribution mainly of the exposed soil and the road in targets with classes that do not belong to them, is due to the fact that, in the segmentation process, the boundaries of blocks presented several "holes", both in the original image and in the image HSI.



Figure 8. Original Image Rating



Figure 9. HSI Image Rating

7. Final Considerations

The extraction of targets in an urban environment is a complex application, due to the heterogeneity of the environment, where it is possible to visualize the details from high resolution images, but they are not enough for the recognition of the patterns that are used to define the limits of the target classes as there are similarities between targets and differences can be subtle. Therefore, the use of auxiliary images helps to discriminate urban targets.

COMPESA uses high resolution images to manually extract information from urban targets and thus generate cartographic material for the company's needs.

The work sought to study the possible contributions of the use of aerial images in the visible range, coming from the PE3D, using indexes and image transformation for the classification of the urban environment. The results showed that: i) the HSI transformation applied to the high resolution image, in addition to highlighting the classes of interest, vegetation, roads, roofs and exposed soil, when stacked maintained the result, presenting a good visibility of the targets ii) the potential the use of VARI and GRVI indices to identify roof-type targets, where these targets were highlighted when the indices were applied; iii) potential of using auxiliary layers together with the images for the extraction of features, generating prominence in the targets of interest; iv) the decision tree classification when applied to the stacked images presents a potential to discriminate the classes.

The original image presented an ease for classification, since it had occurrences of super segmentation, that is, a target was delimited by more than one polygon, which facilitates the training process, allowing more collection of samples for the target of a certain class, while the HSI image presents little segmentation of its targets, which makes the sample collection process and consequently the classification difficult.

The auxiliary layers proved to be viable for the classification of the urban environment, showing that high resolution images need the use of other attributes, such as vegetation indices, in order to better discriminate the targets, in order to improve decision making for generating data. mapping of the urban environment.

Therefore, it is recommended to use other types of auxiliary layers such as LiDAR data to integrate the different attributes and better discriminate targets with similar spectral responses, one of the reasons is that the urban environment has a great

variability of targets. Another way would be to perform target classifications separately, highlighting the homogeneity of blocks and lots, as a form of pre-processing. Since COMPESA is primarily interested in block and plot targets, segmentation needs to reach a level where its segments delimit the targets well.

The methodology used for both images presented satisfactory results, however, it is necessary to carry out more tests that allow the reduction of “noise” in the segmentation stage and consequently a better classification in order to avoid association of classes to which the targets do not belong. The use of the software and the image are beneficial since both are free and do not generate costs.

References

- Blaschke, T. et. Al. “Journal of Photogrammetry and Remote Sensing”, 87 (2014).p. 180-191.
- Korting, T. S. and Fonseca, L. M. G. and Escada, M. I. S. and Câmara, G. “GeoDMA – Um sistema para mineração de dados de sensoriamento remoto”, Anais XIV Simpósio Brasileiro de Sensoriamento Remoto, Natal, Brasil, 25-30 abril, 2009, INPE, p. 7813-7820.
- Meneses, P. R. and Almeida, T. (2012). “Introdução ao processamento de imagens de sensoriamento remoto”.
- Paulo, M. C. M. and Mathia, S. B. B. R. P. and Lacerda, M. and Korting, T. S. and Fonseca, L. M. G. “Comparação dos atributos escolhidos pelo treinamento de classificadores de árvores de decisão com seleção de atributos por filtro”, Revista Brasileira de Cartografia, 2012.
- SILVA, M. L. “Classificação da cobertura do solo urbano utilizando imagens aéreas, produtos LIDAR, índices de vegetação e razão de bandas”. Dissertação (Mestrado no Programa de Pós-Graduação em Ciências Geodésicas e Tecnologias da Geoinformação) – Universidade Federal de Pernambuco, Recife, 2020.
- Witten, I. H. and Frank, E. and Hall, M. A. “Data Mining, Practical Machine Learning Tool and Techniques” 3 ed. 2011.

A tool to predict the growth of urban regions based on QGIS/MOLUSCE using MapBiomass image time series

Mario Arthur Sclafani Pujatti¹, Marconi de Arruda Pereira¹, Leonardo Jose Silvestre²

¹Departamento de Tecnologias (DTECH) – Universidade Federal de São João del Rei (UFSJ)
Ouro Branco – MG – Brasil

²Departamento de Computação e Eletrônica (DCE/UFES)
Universidade Federal do Espírito Santo (UFES) – São Mateus, ES – Brasil

mario.arthur.spujatti@hotmail.com, marconi@ufs.j.edu.br

leonardo.silvestre@ufes.br

Abstract. *This paper presents a method to generate future scenarios of Land-Use and Land-Cover (LULC) classification images by implementing an artificial neural network that can be used to predict urban growth. In this study, LULC data from 1985 to 2020 with annual intervals, obtained through MapBiomass, were used. These data were inserted into a neural network integrated with the MOLUSCE plugin from QGIS to model the possible spatio-temporal changes to simulate the evolution of LULC. MapBiomass is a powerful tool, that uses data from time series from Landsat Satellites and machine learning algorithms to provide reliable products. Our analysis focused on cities that have expanded greatly over the past two decades according to studies made by IBGE. The results obtained were better than those presented in related works, obtaining a kappa value of 0.74 and an accuracy value of at least 80% in all tests performed.*

1. Introduction

Urbanization is one of the most evident global changes. In the last 200 years, while the world population has increased six times, the urban population has multiplied 100 times [Stalker 2000]. In recent decades, with global urbanization, more than half of the world's population lived in cities in 2018, and this proportion is expected to reach 68% by 2050 [UN 2018]. This type of urban expansion and changes in urban land structure has social, economic, ecological and environmental impacts on urban populations and sustainable urban development [Stone et al. 2010]. One of the biggest challenges for global sustainability is the pressure for natural resources driven by intensive urban expansion and economic growth. In this perspective, accelerated urbanization without proper planning is one of the great catalysts contributing to climate change and is an essential factor that compromises ecosystems and their global functionalities [Yang et al. 2020]. As a result, the planning and management in growing urban areas become more complex and difficult. A better understanding of the process of urban growth and the effects of this growth and land use change is required for more efficient planning and management [Leao et al. 2004].

Since the land use and land cover (LULC) change project was launched by the International Geosphere and Biosphere Program (IGBP) and the International Human Dimensions Program (IHDP) on Global Change [Lambin et al. 2000], land use research programs on a global scale have become central to international climate and

environmental change research [Liu et al. 2014]. Thematic maps of LULC serves as a reference for scrutinizing, source administration and forecasting, making it easier to establish plans that balance preservation, competing uses, and growth compressions [Kamaraj and Rangarajan 2022].

On another topic, Artificial Intelligence (AI), has had an increasingly important role in our society in almost every field, including among the territory analysis methods [Gomes et al. 2019]. Different AI systems use different algorithms and are suitable for different purposes [Faceli et al. 2011]. LULC dynamics can be typically modeled by methods that have various implementation complexities and efficiencies [Agarwal 2002] such as artificial neural networks (ANN) [Basse et al. 2014a]. Its applications are suitable for urban and rural environments and it can be applied to different settings. LULC dynamics often result from a complex inter-system combination of factors, a non-trivial collective behavior, that cannot be derived from an individual or a simple collection of systemic analyses [Souza et al. 2022].

Besides being an approach directed towards land usage and cover, we will focus on the urban mesh changes. As so, this work presents a method for predicting and simulating structures of the urban mesh that can contribute to planning and analyzing the structure of the urban mesh. By proposing a way of collecting and reclassifying land usage images, we can build a database of spacial variables that can be used along with data from other sources to simulate the expanse of the urban mesh.

Unlike other approaches, we present a method capable of generating good results using only LULC images that represent the evolution of urban occupation as input. On the other hand, the method also allows the use of the digital model of terrain elevation (MDES) and other parameters to seek to improve the results. Therefore it is possible to achieve results even with only two land cover images. Thus, using the Google Earth Engine (GEE)¹ along with MapBiomass² to get LULC, our proposed method can perform simulations for the entire extension of the national territory, even where getting data would be complex otherwise. Using the MOLUSCE³ interface, it is possible to determine the correlation between the variables used, thus defining their importance for the forecast. Perhaps the most relevant point to justify its use is the replicability of the method, its ready-to-use nature allows users without previous experience to train neural networks to simulate future scenarios.

2. Related Works

Making predictions related to LULC is a frequently revisited area of study. An effective approach has immediate consequences in expanding the use of this type of tool. With the recent advances of studies in the area of machine learning large number of tools that can be applied in the most diverse areas of knowledge are emerging. [Souza et al. 2022] developed a simulation and analysis model of LULC change trends for the year 2036, based on artificial neural networks, in the Chapecó River ecological corridor to provide a robust tool to support decision-making, land use planning and sustainable development.

[Han et al. 2015] shows us an application using a Markov model and a Beijing

¹<https://earthengine.google.com/>

²<https://mapbiomas.org>

³<https://plugins.qgis.org/plugins/molusce/>

case study to describe the related driving factors from land-adaptive variables, regional spatial variables and socio-economic variables and then simulate future land use scenarios from 2010 to 2020. Beijing has undergone rapid urbanization and economic growth since the economic reforms of 1978, and in 2013, the population was approximately 21.15 million. Within the administrative region, the urban population is 18.25 million, which accounts for 86.31% of permanent residents. The new population will need new space and so creates a higher demand for residential land, thus encouraging rapid expansion of the region's urbanized area. According to the overall land utilization plan in Beijing (2006 – 2020), the amount of reserved cultivated land cannot be less than 2147 km^2 by 2020, which means that the amount of cultivated land converted to built-up land is very limited. Thus, an effective application of the proposed method can serve as a powerful tool for planning.

Another practical use can be seen on [Brovelli et al. 2020] work, which proposes a way of monitoring forest cover in the Amazon, using Multi-Temporal Remote Sensing Data and Machine Learning Classification. Emerging countries often suffer from problems related to the modification of basic infrastructure and its balance with the environment. Deforestation causes diverse and profound consequences for the environment and species. Direct or indirect effects can be related to climate change, biodiversity loss, soil erosion, floods, landslides, etc. As such a significant process, timely and continuous monitoring of forest dynamics is important, to constantly follow existing policies and develop new mitigation measures. Using the random forest algorithm as well as GEE a machine learning classification of multispectral satellite imagery on cloud computing service was implemented.

[de Brito et al. 2021] had a similar approach to ours but used some different tools. Also using the MapBiomass database, it aimed to implement a cellular automata-based model using MapBiomass data series as input to develop LULC future scenarios focusing on human activities drivers in a vital watershed for Paraíba state's water security. Using a similar process of imaging and reclassification, our study differs mainly in using MOLUSCE compared to the SIMLANDER chosen by the authors. The effectiveness of our method is evident when the first pre-calibration tests performed by De Brito et al. indicated the accuracy of approximately 56% between the 2015 simulated data and the MapBiomass reference date and after calibration reached an average accuracy of 84.41% for the year 2004, both results lower than those achieved by our process.

3. Methodology

Designing our method required using open source or free services and tools for research purposes. The Brazilian Annual LULC Mapping Project (MapBiomass) is an initiative that involves a collaborative network of biomes, land use, remote sensing, GIS, and computer science experts. The GEE was used together with the MapBiomass project toolkit as a LULC database tool, as well as a way to download them.

We can divide the proposed model for predicting urban growth into three steps. First, the pre-processing step prepares and processes the raw data obtained through MapBiomass, undergoing a reclassification process that is proven to be more effective for the final simulation. The pre-processed data is the model's input. The second step, calibration, has the main goal of establishing the spatial variables for the network training. It

means interpreting which raster types besides LULC have relevant influence to improve the final simulation, as well as discover the optimal values for each parameter used. Finally, the third step intends to validate the results. Thus, Cohen's kappa method is used to compare the generated maps in simulations/predictions with real data from MapBiomass. Figure 1 shows the process.

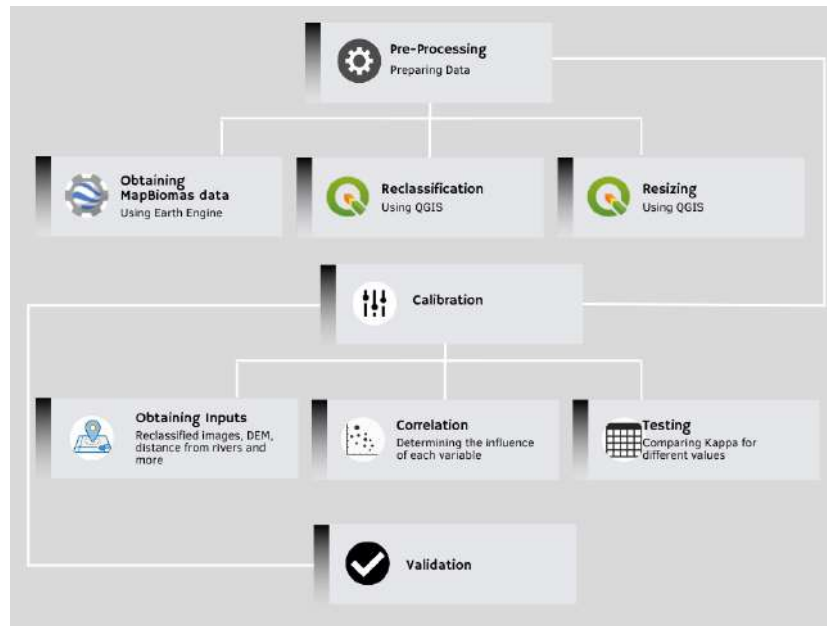


Figure 1. Steps of the proposed method for predicting urban growth

3.1. Pre-processing

Initially, we explored the MapBiomass toolkit for GEE, because it allows visualizing a variety of different land use classifications. The MapBiomass project relies on the GEE platform and its cloud processing and automated classifiers capabilities to generate Brazil's annual land use and land cover time series from 1985 to 2020 (last available collection). GEE combines a multi-petabyte catalog of satellite imagery and geospatial datasets with planetary-scale analysis capabilities. Scientists, researchers and developers use GEE to detect changes, map trends, and quantify differences on the Earth's surface. GEE is now available for commercial use, and remains free for academic and research use. The MapBiomass data uses the time series from Landsat⁴ satellites. The entire process is based on machine learning algorithms providing highly reliable products for the entire territorial extension of the country in a free and accessible way. The spatial resolution of the available satellite images mosaics for each Brazilian biome is 30 m. The mosaics are a composition of pixels in each set of images for a specific period (e.g. filtering the clouds). The periods of the year in which the images are selected vary by region (e.g. wet season in the caatinga biome). We focused on the Brazil subset 6.0 collection. However, MapBiomass' classification method counts almost 50 different classes to represent real-world data as truthfully as possible (Figure 2). As [Congalton 1991] assert, the use of more general categories can be essential when trying to meet predetermined accuracy standards,

⁴<https://landsat.gsfc.nasa.gov/>

and after the first batch of testing, it was confirmed that using LULC with a large number of classes can be harmful to the network learning, and hence to the final simulations.

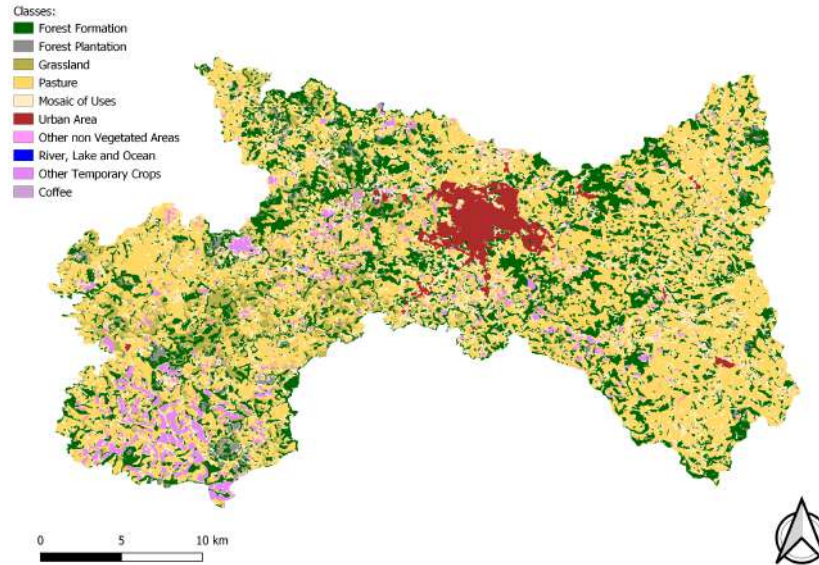


Figure 2. Original Class Distribution

Thereby we recognized the need for a LULC reclassification. Initially, it was considered using only two classes (urban and non-urban), seeing that our analysis is only interested in urban mesh changes, but after the second round of experimentation, it was proven that a too simple reclassification is also damaging on the final simulations.

After the experiments, we concluded that the best option would be to organize the 50 classes into a set of 6 superclasses that successfully encompass the characteristics represented in each of the divisions previously presented. And again, tests were performed showing significant improvement in the proposed simulations. The reclassification process is shown at Figure 3. All forest subtypes were reclassified as forest (ID 1). The same procedure was adopted for the following classes, except for the non-vegetated area: here the urban area continued to be detached from the remaining non-vegetated area. The result of the reclassification process can be seen in Figure 4.

According to Exame⁵ magazine, between 2003 and 2013, medium-sized cities grew faster than any metropolis in the country, so we directed our study to those cities that had the highest population growth in Brazil in this interval, as shown on Table 1. We understand that there is a correlation between population growth and expansion of the urban fabric, and according to studies carried out by [Seto et al. 2011], urban land expansion rates are higher than or equal to urban population growth rates, suggesting that urban growth is becoming more expansive than compact. It allows us to conclude that we made an adequate choice of cities to analyze. It was necessary to use GEE to download the LULC of each city because it allows us to filter the variables according to the area of interest in addition to having methods that can be used to download any

⁵<https://exame.com/brasil/25-cidades-que-sofreram-um-boom-populacional-no-brasil/>

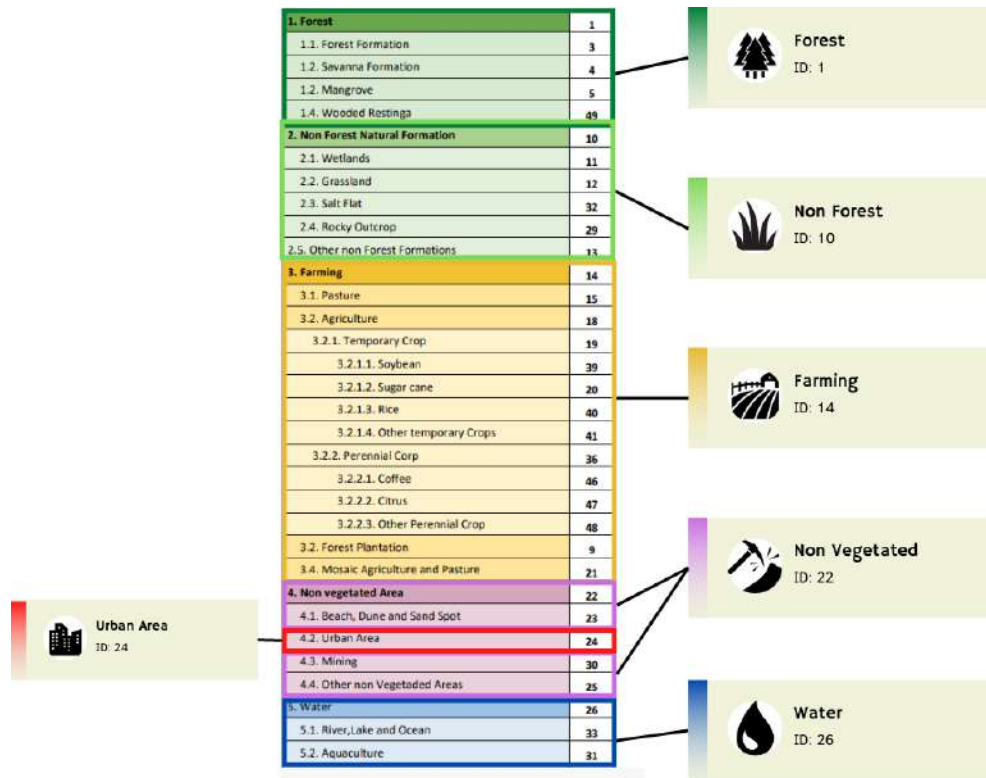


Figure 3. Reclassification Process

variable individually when using the MapBiomias toolkit. So we were able to use these methods to import variables already classified and cropped to Google Drive, enabling the download. Therefore we obtained a vast collection that by itself can potentially be employed for simulating and predicting.

For constructing the land use image database, each of the selected cities had the 36 raw annual images obtained through the MapBiomias toolkit, that were reclassified afterward. However, it was observed in certain cases, a large difference between the extension of the urban mesh and the municipal perimeter, where the perimeter has significantly higher proportions than the urban extension itself. That is, it would be included relatively insignificant data for the study in the analysis. Thus, in these cases, it was necessary to resize the study area, finishing the pre-processing step.

3.2. Calibration

The calibration step consists of analyzing and selecting the variables that will be used for network training. We used QGIS and an add-on called MOLUSCE (Modules for Land Use Change Simulations) to carry out the predictions and simulations. QGIS is open-source, free software. It is a system of geographic information that permits visualizing, editing, and analyzing geo-referenced data, but also executing a variety of helpful algorithms on geoprocessing [Guimarães et al. 2021]. QGIS tools, especially its plugins, are in constant development by a great range of geotechnological sectors. MOLUSCE is one of the community-made plugins available from QGIS menus and provides a set of algo-

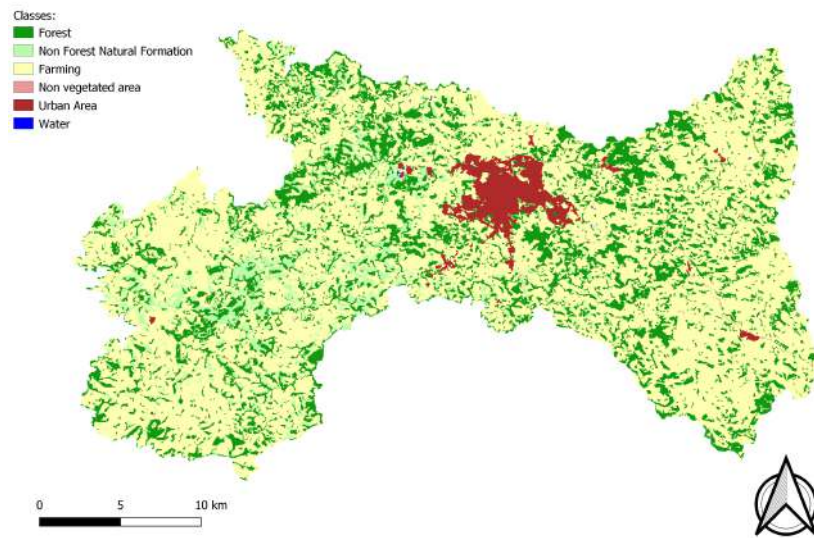


Figure 4. Reclassification Result

	Population Growth	Inicial Population	Final Population
Rio das Ostras	190%	42.024	122.196
Parauapebas	116%	81.428	176.582
Maricá	62%	86.038	139.552
Barcarena	60%	68.604	109.975
Parnamirim	59%	143.598	229.414
Rio Verde	57%	124.753	197.048
Macaé	55%	144.207	275.575
Palmas	49%	172.176	257.903
Lauro de Freitas	45%	127.182	184.383
São José de Ribamar	43%	118.725	170.423

Table 1. Population Growth between 2003 and 2013

rithms for land use change simulations such as neural network, linear regression, among others. MOLUSCE presents a user-friendly interface and has a ready-to-use proposal. It was developed to investigate a range of applications, including studying temporal LULC shifts and projecting future land use, anticipating prospective shifts in land cover and forest cover, and detecting deforestation in sensitive locations[Aneesha Satya et al. 2020].

To choose the model that we would use to perform the simulations, several options were considered. It can be said that every model has its own specialty for addressing the composite issues of LULC. Among these models, cellular automata (CA) are common approaches to simulate LULC and can effectively represent nonlinear spatially stochastic land-use change processes [Batty et al. 1997]. CA are powerful approaches for understanding land-use systems and their integral dynamics [Wu 2002], especially when integrated with other tools, such as ANNs [Basse et al. 2014b]. The Cellular Automata-Artificial Neural Network (CA-ANN) works on what-if scenarios. Therefore, it can

	Year: 1986	Year: 1987	Year: 1988	Year: 1989
Year: 1986	–	0.97522	0.95814	0.94643
Year: 1987	–	–	0.97861	0.95761
Year: 1988	–	–	–	0.97498
Year: 1989	–	–	–	–

Table 2. Pearson's Coefficient

be useful for planning [Araya and Cabral 2010] and land-use change simulation studies [Pahlavani et al. 2017]. The Cellular Automata model employs the transition probabilities from the ANN learning process to describe the LULC changes.

The first step for modeling is to insert the LULCs for the starting date and end of the analysis, in addition to any other spatial variables that you want to use, such as MDES, distance from streets, and other intermediate LULC variables. The plugin compares the geometries of the inserted rasters to get a direct relationship between everything that was entered. Besides correlating the extension and pixel size data, the plugin relies on Pearson's coefficient to simulate a linear correlation between the variables chosen. According to its definition, Pearson's correlation coefficient is the ratio between the covariance of two variables and the product of their standard deviations. Thus, it is essentially a normalized measurement of the covariance, such that the result always has a value between -1 and 1. As the covariance itself, the measure can only reflect a linear correlation of variables and ignores many other types of relationships or correlations. Thus, as illustrated on Table 2, we have a point value that indicates how theoretically relevant each of the variables represents for training and simulation.

After the preliminary tests, we could see that the intermediate LULCs by themselves have a great impact on training, as well as the MDE, the distance from streets, and the distance from rivers. Having calculated Pearson's coefficient, we can proceed to the next step of the process, which is to generate the transition matrix. The transition matrix plays an essential role in analyzing temporal changes within a set of LULC categories. The matrix represents the proportions of pixels changing from one land use category to another. Its rows represent the categories in the initial year, while the columns indicate the same order of LULC categories in the final year. The diagonal entries indicate the size of class stability, and each off-diagonal entry represents the size of the transition from one class to a different class. Values close to 1 in diagonal entries represent the stability of a category. Researchers mostly use transition matrices to compare the temporal changes in different regions.

After creating the matrices, we can generate a map that represents the changes identified in each of the classes from the transition matrix. For each original LULC class, a new class will be generated, representing the change from one state to another. Therefore, we have that in an example of 4 classes we will have 4×4 possible transitions. Each of these transitions receives a code that identifies the change made. The output from the area analysis stage is the change map, indicated on Figure 5, where each pixel is an integer value that indicates its transition.

The next step is to choose the model that will be used to make the forecasts. A CA-ANN model in MOLUSCE is a reliable tool for predicting future LULC that may

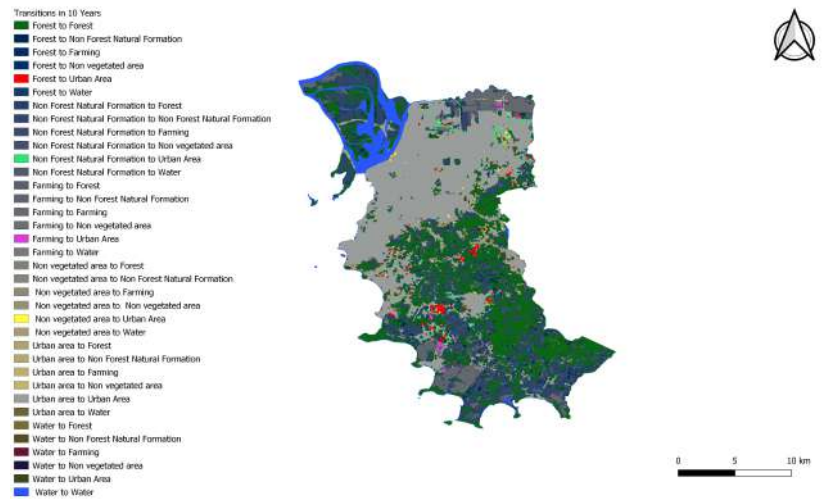


Figure 5. Changes Map

be used in land use planning and management. This approach is being used for predicting the spatial LULC shift because it estimates the pixel's current condition based on its initial situation, adjacent neighborhood eventuality and changeover laws. Moreover, this accurately depicts nonlinear spatial stochastic LULC change processes and produces complex patterns [Saputra and Lee 2019]. The ANN assumes 5 input parameters in addition to the constants that can only be accessed directly in the source code, and their calibration represents an important step to ensure that we can get the best possible result from the simulations. The Neighbourhood variable sets the count of neighboring pixels around the current pixel. Size 1 means 9 pixels (3x3). Learning rate, momentum and max iterations number define learning parameters. The higher the learning rate and momentum, the faster the learning, but the process can be unstable. Hidden Layers takes a list of numbers: $N^1 N^2 \dots N^k$, where N^1 is the number of neurons in the first hidden layer, N^2 in the second, and so on. After several tests, we used the following parameters: Neighbourhood: 5px; Learning Rate: 0.100; Maximum Iterations: 1000; Momentum: 0.050; Hidden Layers: 10.

3.3. Validation

Finally, after training the network, we can generate the simulation. The time interval for the generated simulations is the same interval defined when selecting the initial and final LULC.. That is, if we use a 10 years interval (Initial = 2000 and final = 2010) each iteration of the simulation will also represent an interval of 10 years. The generated LULC maintains the same classes previously represented in the same representation scheme. It is important that we can validate the results of the simulations, and MOLUSCE has the last step that applies the concept of Cohen's kappa, which serves as a point indicator of performance. Cohen's kappa coefficient is a statistic used to measure inter-rater reliability for qualitative items. Is generally considered a more robust measure than the simple percentage agreement calculation, as it takes into account the possibility of agreement occurring by chance. Therefore, a simple and effective way to measure the reliability of the method was to perform the simulations for the year 2020 and compare them with the

	Kappa Coeficient
Barbacena	0.82748
Barcarena	0.83193
Lauro de Freitas	0.78486
Palmas	0.84707
Parauapebas	0.81884
Parnamirim	0.75129
Rio das Ostras	0.86332
Rio Verde	0.85729

Table 3. Simulation Results Validation

maps already produced by MapBiomass, calculating the kappa value for the comparison.

We used an extensive testing method to evaluate what would be the best way to generate the predictions. So, we iterate through pre-defined values for each of the variables, testing all combinations for them. In this way, we evaluated the values that generated a higher kappa with a reasonable training time. Another important point was to repeat the same process for all the cities previously selected, showing that the method is not biased for a specific case.

The proposed method obtained good results in the predictions, as can be seen in Table 3. The lowest kappa achieved by all tests in the 10 cities was 0.75. With the model well established, we can finally generate simulated maps for any future year that is interesting, showing a reliable way to create a basis for geo-economic studies.

To reach one more stage of validation, we applied the proposed method in situations already analyzed by other authors. Then, we compared the results directly, according to the kappa achieved in the validation, which shows similar or better results even with the use of a smaller number of spatial variables. It is evident that the proposed model is effective in simulating future scenarios. Thus, together with the calibration of the algorithm used, we achieved the aim of presenting an easily replicable method of great informational value. A clear example can be observed when we analyze the study carried out by [de Brito et al. 2021] which shows lower Kappa values in all iterations performed, as seen in Table 4.

4. Conclusions

It is undeniable that studies in geo-computing depend on countless variables for an authentic representation of real situations. Even though each forecast relates directly to the spatial variables used, the implementation we used proposes a simple way of obtaining and preparing spatial variables for the entire extension of Brazil.

The proposed process went through several stages of improvement so that the results were the best, considering an implementation simple to be replicated. First, the use of over ten cities in all test steps contributed to the notion that it did not bias the method for certain regions. The experiments carried out concerning the clustering and reclassification of land use were a step to ensure a distribution with enough classes to achieve a satisfactory result. Testing three different classification types, the original with almost 50 classes, a simplistic reclassification with only two classes, and finally, the set

Upper Paraíba River Watershed	Debritto's Kappa	Achieved Kappa
Year: 2005	0.748	0.759
Year: 2006	0.709	0.773
Year: 2007	0.666	0.775
Year: 2008	0.623	0.766
Year: 2009	0.597	0.742
Year: 2010	0.562	0.786
Year: 2011	0.546	0.808
Year: 2012	0.523	0.801
Year: 2013	0.513	0.829
Year: 2014	0.518	0.796
Year: 2015	0.515	0.783
Year: 2016	0.533	0.796
Year: 2017	0.535	0.812
Year: 2018	0.544	0.786

Table 4. Methods Comparison

of 6 superclasses, for all the proposed cities produced a range of results large enough to decide on the best representation to use. A similar method was also used to define the values of the variables used by MOLUSCE to produce a realistic simulation with a viable network training time. In a trial-and-error-like approach, we tested several values for each of the individual variables for later exchanging with each other, thus testing all combinations of pre-selected values.

The tool is available at: https://github.com/marioasp/sim_expansao_urb

Acknowledgements

The authors thank CNPq and FAPEMIG (project APQ-00718-21) for the financial support.

References

- Agarwal, C. (2002). A review and assessment of land-use change models: dynamics of space, time, and human choice.
- Aneesha Satya, B., Shashi, M., and Deva, P. (2020). Future land use land cover scenario simulation using open source gis for the city of warangal, telangana, india. *Applied Geomatics*, 12(3):281–290.
- Araya, Y. H. and Cabral, P. (2010). Analysis and modeling of urban land cover change in setúbal and sesimbra, portugal. *Remote Sensing*, 2(6):1549–1563.
- Basse, R. M., Omrani, H., Charif, O., Gerber, P., and Bodis, K. (2014a). Land use changes modelling using advanced methods: Cellular automata and artificial neural networks. the spatial and explicit representation of land cover dynamics at the cross-border region scale. *Applied Geography*, 53:160 â 171.

- Basse, R. M., Omrani, H., Charif, O., Gerber, P., and Bódis, K. (2014b). Land use changes modelling using advanced methods: Cellular automata and artificial neural networks. the spatial and explicit representation of land cover dynamics at the cross-border region scale. *Applied Geography*, 53:160–171.
- Batty, M., Couclelis, H., and Eichen, M. (1997). Urban systems as cellular automata.
- Brovelli, M. A., Sun, Y., and Yordanov, V. (2020). Monitoring forest change in the amazon using multi-temporal remote sensing data and machine learning classification on google earth engine. *ISPRS International Journal of Geo-Information*, 9(10):580.
- Congalton, R. G. (1991). A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, 37(1):35–46.
- de Brito, H. C., Rufino, I. A. A., and Djordjević, S. (2021). Cellular automata predictive model for man-made environment growth in a brazilian semi-arid watershed. *Environmental Monitoring and Assessment*, 193(6):1–19.
- Faceli, K., Lorena, A. C., Gama, J., and Carvalho, A. C. P. d. L. F. d. (2011). Inteligência artificial: uma abordagem de aprendizado de máquina.
- Gomes, E., Banos, A., Abrantes, P., Rocha, J., Kristensen, S. B. P., and Busck, A. (2019). Agricultural land fragmentation analysis in a peri-urban context: From the past into the future. *Ecological Indicators*, 97:380–388.
- Guimarães, P. V. D., de Arruda Pereira, M., da Costa Teixeira, E. K., Júnior, C. A. D., Arthur, M., and Pujatti, S. (2021). A framework for the generation of the rainwater flow model in streets. XXII Brazilian Symposium on Geoinformatics, GEOINFO2021.
- Han, H., Yang, C., and Song, J. (2015). Scenario simulation and the prediction of land use and land cover change in beijing, china. *Sustainability*, 7(4):4260–4279.
- Kamaraj, M. and Rangarajan, S. (2022). Predicting the future land use and land cover changes for bhavani basin, tamil nadu, india, using qgis molusce plugin. *Environmental Science and Pollution Research*, pages 1–12.
- Lambin, E., Baulies, X., Bockstael, N., et al. (2000). Land-use and land-cover change, implementation-strategy. *IGBP Report No*, 48.
- Leao, S., Bishop, I., and Evans, D. (2004). Simulating urban growth in a developing nation's region using a cellular automata-based model. *Journal of urban planning and development*, 130(3):145–158.
- Liu, J., Kuang, W., Zhang, Z., Xu, X., Qin, Y., Ning, J., Zhou, W., Zhang, S., Li, R., Yan, C., et al. (2014). Spatiotemporal characteristics, patterns, and causes of land-use changes in china since the late 1980s. *Journal of Geographical sciences*, 24(2):195–210.
- Pahlavani, P., Askarian Omran, H., and Bigdeli, B. (2017). A multiple land use change model based on artificial neural network, markov chain, and multi objective land allocation. *Earth Observation and Geomatics Engineering*, 1(2):82–99.
- Saputra, M. H. and Lee, H. S. (2019). Prediction of land use and land cover changes for north sumatra, indonesia, using an artificial-neural-network-based cellular automaton. *Sustainability*, 11(11):3024.

- Seto, K. C., Fragkias, M., Güneralp, B., and Reilly, M. K. (2011). A meta-analysis of global urban land expansion. *PloS one*, 6(8):e23777.
- Souza, J. M. d., Morgado, P., Costa, E. M. d., and Vianna, L. F. d. N. (2022). Modeling of land use and land cover (lulc) change based on artificial neural networks for the chapecó river ecological corridor, santa catarina/brazil. *Sustainability*, 14(7):4038.
- Stalker, P. (2000). *Handbook of the World*. Oxford University Press, USA.
- Stone, B., Hess, J. J., and Frumkin, H. (2010). Urban form and extreme heat events: are sprawling cities more vulnerable to climate change than compact cities? *Environmental health perspectives*, 118(10):1425–1428.
- UN (2018). 68% of the world population projected to live in urban areas by 2050. pages <https://www.un.org/development/desa/en/news/population/2018-revision-of-world-urbanization-prospects.html>. Accessed: 2022-06-26.
- Wu, F. (2002). Calibration of stochastic cellular automata: the application to rural-urban land conversions. *International journal of geographical information science*, 16(8):795–818.
- Yang, J., Gong, J., Tang, W., and Liu, C. (2020). Patch-based cellular automata model of urban growth simulation: Integrating feedback between quantitative composition and spatial configuration. *Computers, Environment and Urban Systems*, 79:101402.

Remote Sensing and Machine Learning on Anomaly Detection at high spectral and temporal dynamics regions in Brazil

Vinícius L. S. Gino, Rogério G. Negri, Felipe N. Souza

¹Instituto de Ciência e Tecnologia (ICT)
Universidade Estadual Paulista “Júlio de Mesquita Filho” (UNESP)
12247-004 – São José dos Campos – SP – Brazil

{vinicius.gino, rogerio.negri, fn.souza}@unesp.br

Abstract. *In climate changes context Remote Sensing tools are widely used and widespread in research. In this sense, Artificial Intelligence rises offering possible improves for environmental monitoring applications using techniques such as Machine Learning for Anomaly Detection applied to Remote Sensing imagery to identify the spatio-temporal changes over the Earth’s surface. This approach is explored in three high dynamic regions in Brazil assessing deforestation, fires and technological disaster areas using One-Class SVM and Isolation Forest methods over MODIS, Landsat and Sentinel images.*

1. Introduction

One of the biggest global challenges is breaking issues like greenhouse gases emission, deforestation, and other disasters impulsed by unstoppable consumption of natural resources [Steffen et al. 2015]. In this sense, the “United Nations 2030 Agenda” provides a multidimensional and holistic vision of this subject, where sustainable development goals rule how to combine human well-being with economic prosperity and environmental protection to guide public policies to mitigate impacts on the environment [Pradhan et al. 2017]. Unfortunately, Brazil lies at the center of debates regarding environmental questions, since it has the largest tropical forest in the world.

The impacts of climate changes are a painful question in the country, once Brazil had the largest area of deforestation between 2010 and 2015 [MacDicken et al. 2016]. The Amazon Forest previously mentioned is influenced by atmospheric and meteorological factors, such El Nino and La Nina whose intensification, combined with human interactions, caused extreme events in past years [Jimenez et al. 2019]. Thus, the biodiversity found in unique biomes such Pantanal is at the mercy of threats caused by land use changes and climate dynamics related with atmospherical air moves from Amazon [Marengo et al. 2021].

In addition, technological disasters are a significant cause of surface dynamics in Brazil. The reason of this approach comes from the recent technological disasters caused by the failures on mining dams in Mariana [do Carmo et al. 2017] and Brumadinho [Rotta et al. 2020], which resulted in the death of hundreds of people in addition to significant environmental impacts. Face to these events, the development of strategies and tools to analyze and monitor mining dams has demanded attention.

In this scenario, Remote Sensing technology rises as a convenient tool for observing and analyzing the Earth’s surface. Beyond allowing register the information in differ-

ent spectral wavelengths, the remote sensors also allow wide spatial and temporal analysis [Jensen 2009]. Additionally to Remote Sensing data, the Machine Learning techniques encompass the construction of algorithms able to identify and extract information from large bases of data, which includes diverse studies and applications with Remote Sensing data [Lary et al. 2016]. Anomaly Detection comprises a kind of unsupervised Machine Learning technique that may be applied in Remote Sensing data to automatically identify the temporal changes and dynamics over the Earth’s surface [Guo et al. 2016].

In the light of the presented discussions, this study addresses the use of Anomaly Detection and Remote Sensing data to identify regions with high spectral-temporal dynamics. Furthermore, this research proposes and implements a prototype of an “anomaly monitoring and warning system” fed by images acquired by the MODIS, Sentinel and Landsat programs/satellites. Functionalities of the Google Earth Engine platform support such implementation. A study case focuses on analyzing the regions at the center of environmental debate such the city of Altamira, Para, in Amazon Forest and areas affected by the dams collapse in Mariana and Brumadinho.

2. Theory background

2.1. Anomaly Detection

Among the different techniques that permeate Machine Learning, Anomaly Detection identifies events/elements with significantly distinct behavior compared to other observations. Usually, such techniques have been used in the identification of bank fraud, checking for intruders in security systems, and in supporting medical analysis [Gu et al. 2019]. In addition to these applications, anomaly detection techniques are highlighted as a potential tool for the environmental monitoring [Dereszynski and Dietterich 2011].

The Breaks For Additive Season and Trend (BFAST) [Lambert et al. 2013], Local Outlier Factor (LOF) [Ma et al. 2013], Elliptic Envelope [Hoyle et al. 2015] and One-Class Support Vector Machine (OC-SVM) [Chen et al. 2001] and Isolation Forest (IF) [Liu et al. 2008] are example of Anomaly Detection methods found in the literature. In special, the two latter mentioned methods have been successfully employed in remote sensing studies [Rembold et al. 2013, Holloway and Mengersen 2018].

As a variant of the well-known and attractive Support Vector Machine (SVM) method, the OC-SVM [Chen et al. 2001] deals with quantile estimation and anomaly detection problems. Conceptually, starting from a set of observations $\mathcal{Z}$, the OC-SVM method provides a model capable of classifying the objects as part of a set of non-anomalous elements according to a probability ν of false-positive occurrence.

It is worth noting that the OC-SVM is parameterized by $\nu \in [0, 1]$ and other parameters related to the adopted kernel function. Further details on kernel functions are discussed in [Shawe-Taylor et al. 2004].

The Isolation Forest (IF) [Liu et al. 2008] comprises a low-computational cost method able to overcome the difficulties when dealing with large databases. This method has been used in Remote Sensing studies [Li et al. 2019] and other analyses involving digital image processing [Alonso-Sarria et al. 2019].

In summary, the IF embodies an ensemble of decision trees, in this case, called “isolated tree” (IT). According to the conceptual idea behind this method, when the

data/objects are submitted to classification in a decision tree scheme, the anomalies tend to present a short path to the root node. The expected length of this path is strictly dependent on the number of decision trees in the ensemble and the size of the dataset [Lesouple et al. 2021].

The definition of an IT starts from a sample set $\{\mathbf{x}_1, \dots, \mathbf{x}_m\}$, where $\mathbf{x}_i = [x_{i1}, \dots, x_{id}]^T \in \mathbb{R}^d$ with components express a specific attribute in m observations. This dataset may also be represented as a matrix $\mathbf{X}$ whose columns are the vectors $\mathbf{x}_i$, for $i = 1, \dots, m$. The nodes of a IT may be either internal or external. While the earlier have two descendants, the external node has no descendent and are called “leaf”. Based in this structure, the IT sequentially randomly select a value p in the q -th attribute to split $\mathbf{X}$ into two descendants. After recursively perform this process, the IT is defined. As stop criterion for the IT expansion, is assumed: (i) the IT reaches its length limit; (ii) $|\mathbf{X}| = 1$; or (iii) all the columns of $\mathbf{X}$ are equal.

Regarding the IT structure, the Anomaly Detection process is performed by scores assigned to each $\mathbf{x}_i$ according to the root-to-leaf path length that such vector pass-through the IT, represented by $h(\mathbf{x}_i)$. The average estimate of $h(\mathbf{x}_i)$ for the external nodes is the same as an unsuccessful search in a Binary Search Tree, expressed as:

$$c(m) = 2H(m-1) - \frac{2(m-1)}{m} \quad (1)$$

where $H(i) = \ln(i) + 0.5772156649$ is a harmonic number [Havil 2003] and $c(m)$ is the average estimate of $h(\cdot)$ considering the m observations. In turn, the anomaly score is:

$$s(\mathbf{x}_i, m) = 2^{-\left(\frac{E(h(\mathbf{x}_i))}{c(m)}\right)} \quad (2)$$

where $E(h(\mathbf{x}_i)) = \frac{1}{q} \sum_{i=1}^q h(\mathbf{x}_i)$ is the mean of $h(\mathbf{x}_i)$ from a collection of ITs.

Therefore, it can be inferred that if $E(h(\mathbf{x}_i))$ tends to zero, the score tends to 1, representing then an anomaly. On the other hand, when $h(\mathbf{x}_i)$ tends to $m-1$, s tends to 0, showing very likely regular data. Furthermore, when $E(h(\mathbf{x}_i))$ tends to $c(m)$, $s(\mathbf{x}_i, m)$ tends to 0.5 and then there is no anomaly distinction.

2.2. Spectral Indices

A spectral index comprises a combination of two or more spectral bands to provide a particular representation of the Earth’s surface. Among a plethora of spectral indices proposed in the literature, the vegetation indices take into account the spectral response of chlorophyll targets concerning electromagnetic radiation from the Sun [Moreira 2000].

One of the most used vegetation indices for canopy characterization is the Normalized Difference Vegetation Index (NDVI) [Rouse et al. 1974], which uses the red and infrared bands as input data. This index has various application purposes, for example, monitoring and mapping crops, droughts, pest damage, agricultural productivity, hydrological modeling, and others [Xue and Su 2017].

The Normalized Difference Water Index (NDWI) [Gao 1996] comprises a spectral index based on the region of electromagnetic spectrum sensitive to water presence. Its use allows detecting particulate matter and suspended sediments in water columns.

Let consider $\mathcal{I}(s) = \mathbf{x}$ where the components x_{Green} , x_{Red} and x_{NIR} stands for the radiometric response at the green, red and near-infrared wavelengths. The NDVI and NDWI values at the position s is computed by $\frac{x_{NIR} - x_{Red}}{x_{NIR} + x_{Red}}$ and $\frac{x_{Green} - x_{NIR}}{x_{Green} + x_{NIR}}$, respectively.

3. Proposal of multitemporal anomaly detection

3.1. Conceptual formalization

Figure 1 depicts a general overview of the proposed method for multitemporal anomaly detection.

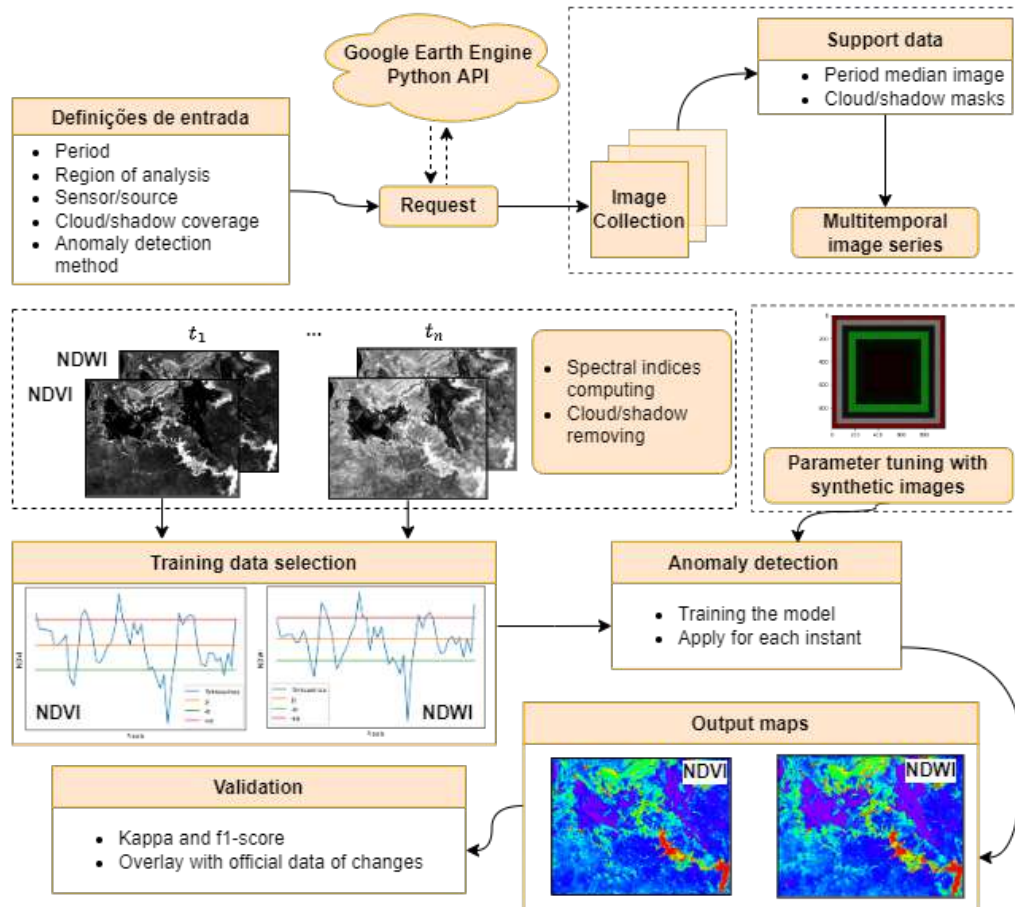


Figure 1. Overview of the proposed method.

Accordingly to this structure, as an initial step, it is defined the period of analysis, the region of interest, a cloud cover threshold, and a remote sensor as a data source. The anomaly detection method is also defined in the initial step. Such configuration (except the anomaly detection model) is submitted as a request to the Google Earth Engine (GEE), which consequently returns a collection of images that gives place to a multitemporal image series. A median image and cloud/shadow cover masks are determined from such image series as support data for posterior use. In a second stage, the NDVI

and NDWI are computed at each instant and then subtracted from the median image of period for that study area to translate all the data around a common central tendency (i.e., the zero). Moreover, information from areas affected by cloud and shadow occurrences are disregarded after applying the previously defined masks. After, the NDVI and NDWI translated values in $[-\alpha\sigma, +\alpha\sigma]$ are used to train an anomaly detection model F and classify the complete dataset. The σ is the dataset standard deviation and $\alpha \in \mathbb{R}$ is an adopted scale factor. Lastly, a map about the multitemporal dynamics is produced according to an anomaly counting over the analyzed period. Also, a map of p -value based on the “run test of randomness” [Siegel and Castellan 1988] allows identifying regions with high confidence regarding the occurrence of the changes.

3.2. Implementation details

The Python 3.8 was the *programming language* adopted to implement the proposed method, as the monitoring prototype. Additionally, the *Scikit-Learn library* was used to apply the Anomaly Detection methods. Moreover, the *Pandas library* was employed to organize the information.

The *Anomaly detection models* are trained with basis on observed values of a previously defined spectral index (i.e., NDVI or NDWI) in $[-\alpha\sigma, +\alpha\sigma]$, where σ is the standard deviation of considered spectral index and $\alpha = 0.5$ is a constant adopted to control the training set regularity.

Lastly, the *Google Earth Engine (GEE) Application Programming Interface (API)* is used to access the Remote Sensing image catalogs and obtain the multitemporal image series according to the defined period, region, and sensor, based in Python. Landsat and Sentinel data are considered in this study. The cloud occurrence threshold of 20% inside the region of analysis is admitted to disregarding useless scenes.

4. Experiments

4.1. Synthetic data

To validate the purpose and optimize previously mentioned Anomaly Detection methods from Remote Sensing applications, this project is assessed using 100 synthetic images segmented into six degrees of anomaly probability for equal areas.

Once defined the image series, different parameter configurations were tested by a grid-search procedure with 10-fold cross-validation and the respective results were assessed in terms of F1-Score [Rijsbergen 1979] measure by considering six “classes of dynamic”. Such classes stands form “no anomalies”, “very low”, “low”, “medium”, “high” and “very high” when the anomaly frequencies are 0%, 1–20%, 21–40%, 41–60%, 61–80% and 81–100%, respectively.

4.2. Study area and Remote Sensing data

In order to assess the method proposed and discussed at Section 3, it is carried a practical application regarding the analysis of temporal dynamics in regions divided into 2 groups: Mariana-MG and Brumadinho-MG affected after the respective dam collapses; Altamira-PA one of largests cities of Brazil in territorial area whose vegetation are threat by human exploration and climate changes. Figure 2 shows the area locations.

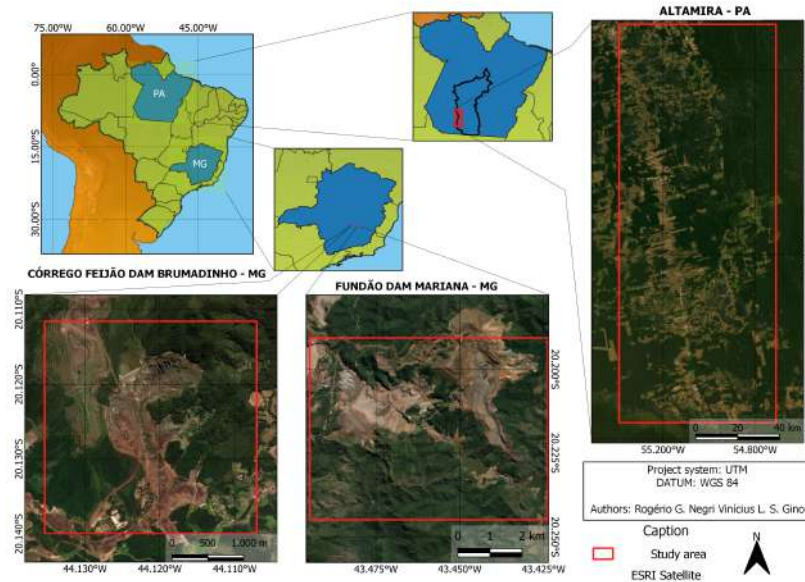


Figure 2. Spatial location of study areas.

It is worth highlighting that the Mariana (Fundão) and Brumadinho (Córrego do Feijão Mine I) dams are located in Minas Gerais (MG). These areas are considered strategic for the development of mining activity in Brazil, a sector responsible for 4% of the national GDP and the generation of more than 2 million indirect jobs [IBRAM 2020]. The disruption of these structures impacted the surrounding landscape, initially surrounded by vegetation characteristic of the Atlantic Forest biome. Moreover, these dams were built following the upstream heightening, which is less costly but with the greater risk of disruptions [Thomé and Passini 2018].

In turn, Altamira is located in the state of Pará, north region of Brazil, inside Amazon Forest. Based in Brazilian Amazon Deforestation Monitoring Program (PRODES) data, since 2015 the deforestation rate increases in Amazon biome caused by many factors that reach that region, such agricultural frontier dynamics, land grabbing activity and illegal mining, being a key point to climate changes intensification [Silva Junior et al. 2021].

In this sense, Remote Sensing data used in this framework considered the area of dams structure and detailing capability given by each sensor: MODIS, Landsat-8 OLI and Sentinel MSI. Table 4.2 show the sensor used for each study area, the period of analysis and the number of images collected in GEE API considering a threshold in cloud occurrence of 10% for each instant.

Area	Sensor	Resolution [m]	Period	Images
Altamira	MODIS	250	2010-2021	278
Brumadinho	Sentinel MSI	10	2016-2021	67
Mariana	Landsat-8 OLI	30	2013-2021	79

Table 1. Data information from each study area.

4.3. Results and discussion

Using synthetic data, Figure 3(b) displays the f1-scores comparison assessed to define the Anomaly Detection Methods parameters and table 2 shows them.

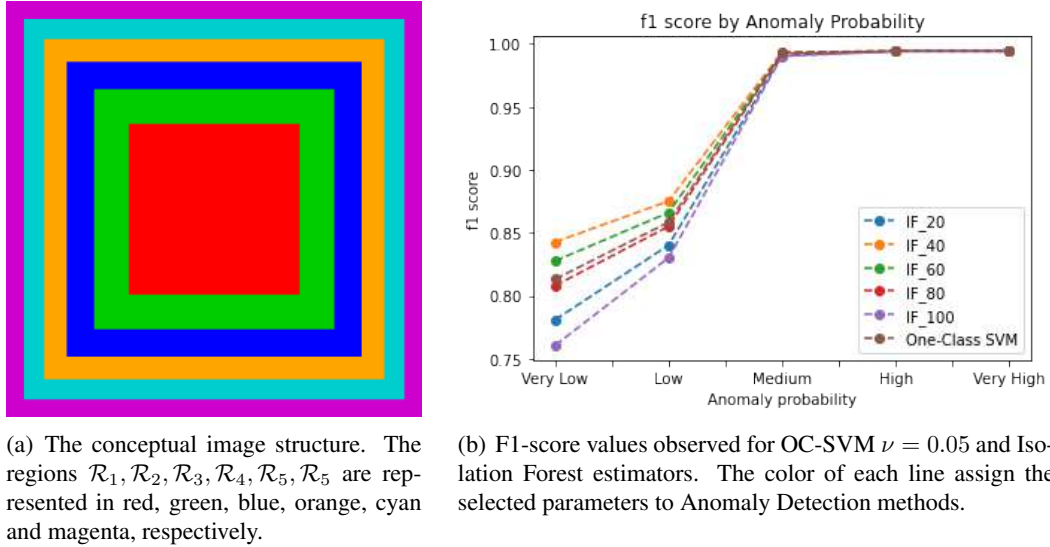


Figure 3. Synthetic data composition and parameter optimization.

OC-SVM		Isolation Forest	
Parameter	Value	Parameter	Value
kernel	rbf	number of estimators	40
kernel coefficient (γ)	auto	max samples, contamination, bootstrap, verbose, warm start	default
upper bound on the fraction of training errors (ν)	0.05		
tolerance, shrinking, cache size, verbose, max iterations	default		

Table 2. Anomaly Detection Methods parameters.

Figures 4, 5 and 7 depicts a bi-temporal comparison using color compositions and the respective multitemporal dynamic maps in terms of “anomaly detection counting” and “ p -value”. The first one was obtained by percentage discretization of anomaly detection data, where the lowest 20% represents “Very low” label, and so on. The NDVI and NDWI values are considered to obtain the results for Brumadinho, Altamira and Mariana areas, respectively.

To validate and compare the results generated by the proposed method, reference samples collected from change maps of moments before and after dam failures were divided into (i) No change areas; (ii) Change areas. These samples were applied at Anomaly Detection maps, which can be observed in the histograms highlighted by Figure 6, whose expected results were the decrease of “No changes” bars as they increase “Changes” bars along anomaly count axis.

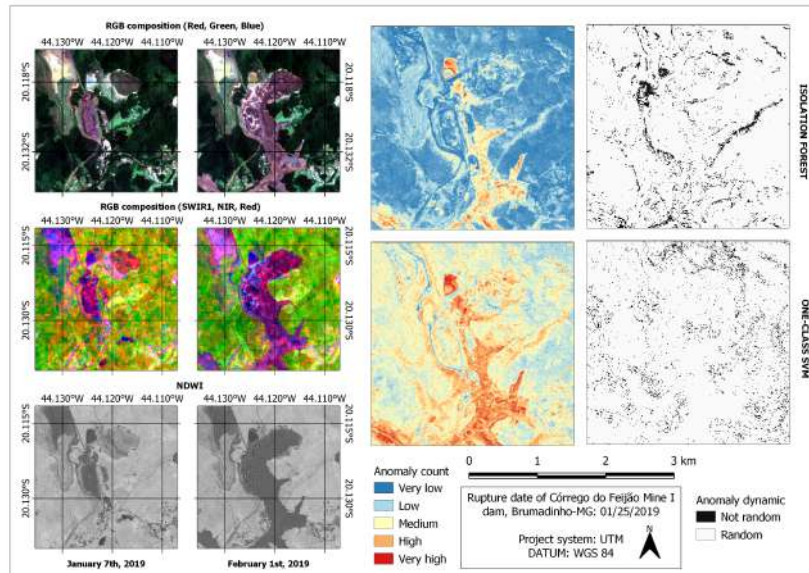


Figure 4. Results using NDWI for Brumadinho dam area.

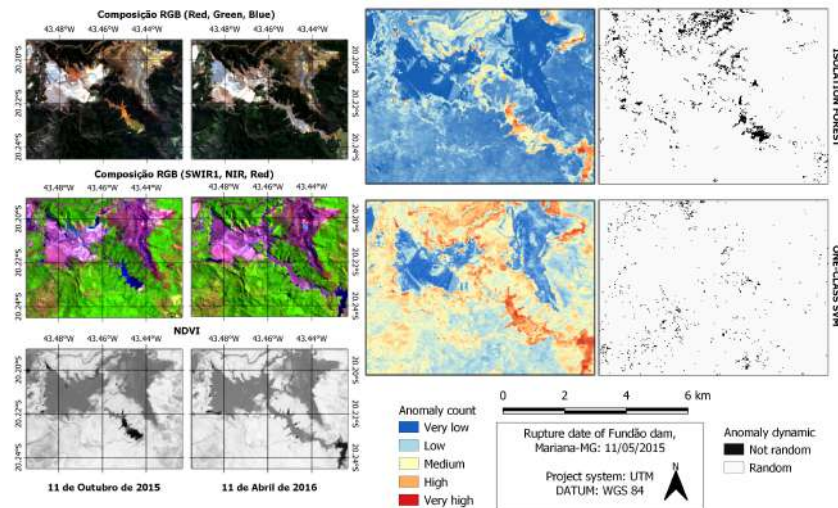


Figure 5. Results using NDVI for Mariana dam area.

Concerning that results, the validation is closed with a build of confusion matrix where x axis is represented by changed and unchanged regions in time series, in turn in the y axis represents the samples of changes and no changes. The anomaly count data, exposed in five labels, was clustered in a binary classification, were 'Medium' count of anomalies was interpreted as transition areas. 'Very Low' and 'Low' labels were grouped as 'Unchanged' while 'High' and 'Very High' labels as 'Changed'. Table 3 shows the validation metrics *Kappa Score* and *Accuracy*.

For assessing Altamira results, PRODES deforestation data between 2010 and 2021 was considered and overlaid under anomaly detection counting maps by three

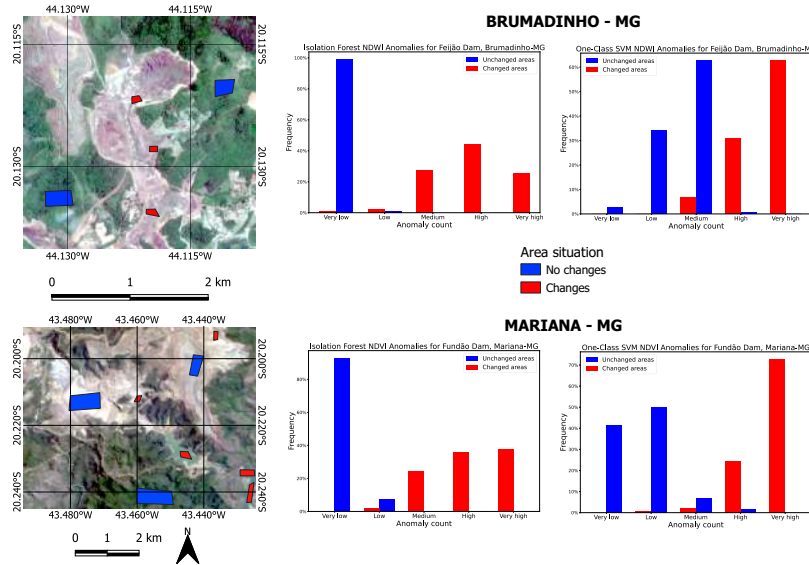


Figure 6. Anomaly count for reference samples and change map comparison.

Study area	Index	Method	Accuracy	<i>Kappa</i>
Brumadinho	NDWI	IF	0.849	0.747
		OC-SVM	0.651	0.316
Mariana	NDVI	IF	0.867	0.775
		OC-SVM	0.943	0.888

Table 3. Validation metrics for each method and study area

year periods through previously mentioned dates. Figure 7 shows the Anomaly Detection maps and validation with PRODES data.

Focusing on the “anomaly detection counting” maps, it is possible to observe that while the IF method delivers more consistent results, OC-SVM tends to overestimate the frequency of anomaly occurrence, evidenced in validation of dam areas and in agricultural fields in Altamira. Low-dynamic regions, like vegetation and exposed soil, are also highlighted when the proposed method is equipped with the IF model.

Regarding the p -value maps, under a 5% significance, the pixels in black represents regions with not random behavior in terms of anomaly/regular occurrence over time. Consequently, such regions demand attention when analyzing the obtained maps. Among the possible causes are seasonal changes showed by targets like water bodies and vegetation. In general, the p -value mapping results achieved with the IF model are more consistent than those using the OC-SVM.

The whole process involved considerable computational costs. The reference machine was a desktop with 16 GB RAM and 500 GB of SSD memory. Generally, the time of image processing and application of Anomaly Detection Methods can vary between 1.5 and 3 hours depending on spatial resolution of Remote Sensing data and study area dimensions.

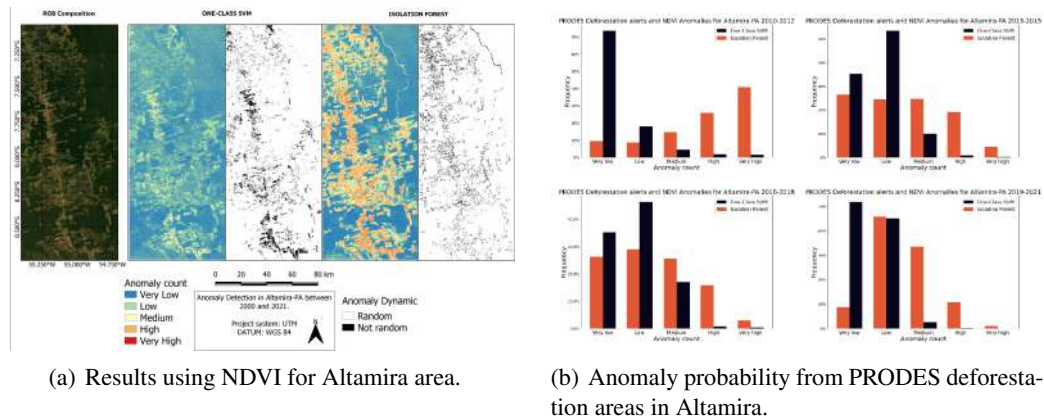


Figure 7. Anomaly Detection results and PRODES overlay from Altamira region.

5. Conclusions

Based on the presented results, it is possible to verify that the proposed method, viewed as an environmental monitoring system prototype, could identify anomalies that correspond to targets with high spectral-temporal dynamics.

It is noteworthy that the assessed Anomaly Detection models have different precision. The IF method was able to distinguish with better contrast the regions of anomalies and regular and provide more consistent p -value maps (useful to identify seasonal changes). OC-SVM method, in turn, was more sensible to change detection often classifying unchanged regions as anomalies.

In future works could be addressed the suppression of seasonal trends and the combination of multiple spectral indices to improve the proposed prototype. Furthermore, the expansion of study areas and availability of this approach and availability of results for the general public in easy visualization and manipulation tools could be explored.

Acknowledgments

The authors thank FAPESP (grant 2018/01033-3, 2020/14664-1 and 2021/01305-6) and CNPq for their financial support of this research.

References

- Alonso-Sarria, F., Valdivieso-Ros, C., and Gomariz-Castillo, F. (2019). Isolation forests to evaluate class separability and the representativeness of training and validation areas in land cover classification. *Remote Sensing*, 11(24):3000.
- Chen, Y., Zhou, X. S., and Huang, T. S. (2001). One-class svm for learning in image retrieval. In *Proceedings 2001 International Conference on Image Processing (Cat. No. 01CH37205)*, volume 1, pages 34–37. IEEE.
- Dereszynski, E. W. and Dietterich, T. G. (2011). Spatiotemporal models for data-anomaly detection in dynamic environmental monitoring campaigns. *ACM Transactions on Sensor Networks (TOSN)*, 8(1):1–36.

- do Carmo, F. F., Kamino, L. H. Y., Junior, R. T., de Campos, I. C., do Carmo, F. F., Silvino, G., Mauro, M. L., Rodrigues, N. U. A., de Souza Miranda, M. P., Pinto, C. E. F., et al. (2017). Fundão tailings dam failures: the environment tragedy of the largest technological disaster of brazilian mining in global context. *Perspectives in ecology and conservation*, 15(3):145–151.
- Gao, B.-C. (1996). NDWI – A normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266.
- Gu, J., Wang, L., Wang, H., and Wang, S. (2019). A novel approach to intrusion detection using svm ensemble with feature augmentation. *Computers & Security*, 86:53–62.
- Guo, Q., Pu, R., and Cheng, J. (2016). Anomaly detection from hyperspectral remote sensing imagery. *Geosciences*, 6(4):56.
- Havil, J. (2003). Gamma: exploring euler’s constant. *The Australian Mathematical Society*, page 250.
- Holloway, J. and Mengersen, K. (2018). Statistical machine learning methods and remote sensing for sustainable development goals: A review. *Remote Sensing*, 10(9):1365.
- Hoyle, B., Rau, M. M., Paech, K., Bonnett, C., Seitz, S., and Weller, J. (2015). Anomaly detection for machine learning redshifts applied to sdss galaxies. *Monthly Notices of the Royal Astronomical Society*, 452(4):4183–4194.
- IBRAM (2020). Informações sobre a economia mineral brasileira 2020. Technical report, Instituto Brasileiro de Mineração.
- Jensen, J. R. (2009). *Remote sensing of the environment: An earth resource perspective*. Pearson Education India, 2 edition.
- Jimenez, J. C., Marengo, J. A., Alves, L. M., Sulca, J. C., Takahashi, K., Ferrett, S., and Collins, M. (2019). The role of enso flavours and tna on recent droughts over amazon forests and the northeast brazil region. *International Journal of Climatology*.
- Lambert, J., Drenou, C., Denux, J.-P., Balent, G., and Cheret, V. (2013). Monitoring forest decline through remote sensing time series analysis. *GIScience & Remote Sensing*, 50(4):437–457.
- Lary, D. J., Alavi, A. H., Gandomi, A. H., and Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*, 7(1):3–10.
- Lesouple, J., Baudoin, C., Spigai, M., and Tournet, J.-Y. (2021). Generalized isolation forest for anomaly detection. *Pattern Recognition Letters*, 149:109–119.
- Li, S., Zhang, K., Duan, P., and Kang, X. (2019). Hyperspectral anomaly detection with kernel isolation forest. *IEEE Transactions on Geoscience and Remote Sensing*, 58(1):319–329.
- Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 eighth ieee international conference on data mining*, pages 413–422. IEEE.
- Ma, H., Hu, Y., and Shi, H. (2013). Fault detection and identification based on the neighborhood standardized local outlier factor method. *Industrial & Engineering Chemistry Research*, 52(6):2389–2402.

- MacDicken, K., Jonsson, Ö., Piña, L., Mauro, S., Contessa, V., Adikari, Y., Garzuglia, M., Lindquist, E., Reams, G., and DâAnnunzio, R. (2016). Global forest resources assessment 2015: how are the world's forests changing?
- Marengo, J. A., Cunha, A. P., Cuartas, L. A., Deusdará Leal, K. R., Broedel, E., Seluchi, M. E., Michelin, C. M., De Praga Baião, C. F., Chuchón Ângulo, E., Almeida, E. K., et al. (2021). Extreme drought in the brazilian pantanal in 2019–2020: Characterization, causes, and impacts. *Frontiers in Water*, 3:13.
- Moreira, R. d. C. (2000). Influência do posicionamento e da largura de bandas de sensores remotos e dos efeitos atmosféricos na determinação de índices de vegetação. *São José dos Campos. 181p. Dissertação (Mestrado em Sensoriamento Remoto)-INPE.*
- Pradhan, P., Costa, L., Rybski, D., Lucht, W., and Kropp, J. P. (2017). A systematic study of sustainable development goal (sdg) interactions. *Earth's Future*, 5(11):1169–1179.
- Rembold, F., Atzberger, C., Savin, I., and Rojas, O. (2013). Using low resolution satellite imagery for yield prediction and yield anomaly detection. *Remote Sensing*, 5(4):1704–1733.
- Rijsbergen, C. J. V. (1979). *Information Retrieval*. Butterworth-Heinemann, USA, 2nd edition.
- Rotta, L. H. S., Alcantara, E., Park, E., Negri, R. G., Lin, Y. N., Bernardo, N., Mendes, T. S. G., and Souza Filho, C. R. (2020). The 2019 brumadinho tailings dam collapse: Possible cause and impacts of the worst human and environmental disaster in brazil. *International Journal of Applied Earth Observation and Geoinformation*, 90:102119.
- Rouse, J. W., Haas, R. H., Schell, J. A., Deering, D. W., et al. (1974). Monitoring vegetation systems in the great plains with erts. *NASA special publication*, 351(1974):309.
- Shawe-Taylor, J., Cristianini, N., et al. (2004). *Kernel methods for pattern analysis*. Cambridge university press.
- Siegel, S. and Castellan, N. (1988). *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill international editions statistics series. McGraw-Hill.
- Silva Junior, C. H., Pessoa, A., Carvalho, N. S., Reis, J. B., Anderson, L. O., and Aragão, L. E. (2021). The brazilian amazon deforestation rate in 2020 is the greatest of the decade. *Nature Ecology & Evolution*, 5(2):144–145.
- Steffen, W., Broadgate, W., Deutsch, L., Gaffney, O., and Ludwig, C. (2015). The trajectory of the anthropocene: the great acceleration. *The Anthropocene Review*, 2(1):81–98.
- Thomé, R. and Passini, M. L. (2018). Barragens de rejeitos de mineração: características do método de alteamento para montante que fundamentaram a suspensão de sua utilização em minas gerais. *Ciências Sociais Aplicadas em Revista*, 18(34):49–65.
- Xue, J. and Su, B. (2017). Significant remote sensing vegetation indices: A review of developments and applications. *J. Sensors*, 2017:1353691:1–1353691:17.

Characterization of Fracture Geometry with Computed Microtomography Rock Images

Letícia da Silva Bomfim¹, Guilherme Daniel Avansi²
Alexandre Campana Vidal³, Helio Pedrini¹

¹Institute of Computing, University of Campinas (UNICAMP)
Campinas, SP, Brazil, 13083-852

²SOLPE - Oil Reservoir Solutions
Campinas, SP, Brazil, 13083-898

³Institute of Geosciences, University of Campinas (UNICAMP)
Campinas, SP, Brazil, 13083-855

leticiaabomfim4@gmail.com, guiavansi@gmail.com,
helio@ic.unicamp.br, vidal@unicamp.br

Abstract. *Fractures are planar structures that can be found inside a rock, in which there is the ability to accumulate and transport fluids. From this important structure, we developed a computational framework that analyzes computed tomographic images by extracting their geometric characteristics. For this analysis, image processing techniques were developed that seek to identify the structure of the contours obtained through data segmentation. Thus, this work presents the techniques and algorithms that were implemented for the evaluation of these parameters, with the main objective of helping the reservoir area and fluid characterization, avoiding destructive analysis of the samples and reducing the uncertainties inherent to the characterization of the productive life of a field.*

1. Introduction

Many petroleum, geothermal and water supply reservoirs are formed in fractured rocks [Council 1996]. This provides a great importance to the study of internal structures, such as pores and fractures, found in rocks, because, through them, paths are formed that provide the propagation and conduction of these chemical substances.

Fractures are the planes along which stress has caused partial loss of cohesion in the rock. It is a relatively smooth flat surface, representing plane of weakness (discontinuity) [Singhal and Gupta 2010], and its existence can cause or prevent a flow of fluid between them. Furthermore, as reported by Li et al. [Li et al. 2021], stress-dependent fracture permeability also brings implications to oil and gas production, underground CO₂ storage, and deep waste disposal containment.

Thus, the determination of their geometry and properties is of paramount importance to feed flow simulators, which use mathematical models of flow in porous media to describe the physical phenomenon of mass and/or energy flow, based on macroscopic parameters. Zhu et al. [Zhu et al. 2021] investigated the impact of the geometric properties

of stochastic fracture networks (fracture lengths, orientations, openings and fracture center positions), showing the great implication of these parameters in the main macroscale flow properties.

Thus, to analyze the material of a reservoir, it is necessary to collect a rock fragment that allows the study of the macroscopic parameters that represent it. However, the process of collecting this material is extremely arduous and costly, making methods that do not harm the integrity of the sample the best way to preserve and perpetuate the material collected.

Another point that reinforces the need for non-destructive methods is the difficulty of analyzing the sample to obtain its properties. Jing and Stephansson [Jing and Stephansson 2007] stated that an explicit representation of all individual fractures can be obtained directly to generate a deterministic model for analysis or design, but that this process, when done manually, is impractical due to the large number of fractures with different distributions.

The computed microtomography technique has been used to prevent the periporous characteristics from being modified as laboratory experiments are carried out, favoring the durability and originality of the sample, in addition to allowing a deeper analysis than a study manual offers, since image processing techniques can be applied to the images obtained. It did not come to replace laboratory experiments but to complement them. As an example, Figure 1 shows a rock sample with the presence of fracture.

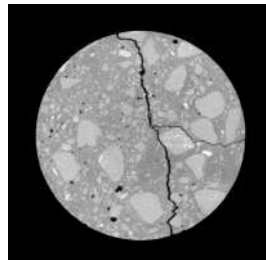


Figure 1. MicroCT image of carbonate rock with the presence of a fracture.
Source: [du Plasis 2014].

Based on this, from the group of tomographic images generated through a sample, we developed an algorithm that acts in a two-dimensional (2D) way in its processing, aiming to identify the fractures present, and characterize them according to their geometry. This process is initially carried out by segmentation, where the structures are extracted and a mask image is also generated that delimits the area of interest of the rock, allowing the analysis of the detected components.

In addition to analyzing each fracture individually, a general analysis of the 2D slice being processed is also performed, in which the data regarding the complete set of fractures found are analyzed in order to quantify them in number, density, and percentages related to slope and shape.

This paper is organized as follows. In Section 2, a theoretical framework presents relevant concepts about the characteristics to be extracted from fractures and pores. In Section 3, the algorithm development methodology and its interface are described. In

Section 4, the experimental results are presented and discussed. Finally, conclusions and directions for future work are described in Section 5.

2. Background

In this section, we contextualize recent works in the area of rock analysis from computed microtomography and then we address the characteristics that will be evaluated.

2.1. Related Work

The analysis of carbonate rocks is a very common practice in the study of fluid flow [Santos et al. 2022]. This is due to the fact that its properties can guide the exploration of reservoirs through the characteristics obtained from its main components, the pores and fractures. One of the oldest ways, and still used due to its precision, for the analysis of this material is the injection of non-wetting fluids. According to Lucia [Lucia et al. 2003], the most accurate method is helium gas expansion where the pressure difference will indicate the pore volume. In this line of study, Al Shafloot et al. [Al Shafloot et al. 2021] and Yang et al. [Yang et al. 2021] also developed a study that investigates fracturing behavior through the use of non-wetting fluids, such as carbon dioxide and nitrogen, making use of continuous monitoring of in situ details using X-ray Computed Tomography (CT).

Basan et al. [Basan et al. 1997] carried out a study using three different techniques for data analysis on a core material: capillary mercury injection pressure (MICP), backscattered electron imaging (BSEI) and nuclear magnetic resonance (NMR). In their work, the authors concluded that the three techniques provide a data distribution suitable for cross-correlation in order to obtain a visual and quantitative analysis of the pore structure. They also mentioned that NMR techniques have the potential not only to impact the quality of reservoir simulations, but also to provide important insights into what regulates reservoir performance.

Thus, with the identification of the potential for data acquisition through X-ray techniques, several works have been developed to study rock samples in a deeper way, taking advantage of the high capacity to visualize and study the microstructures. Cnudde and Boon [Cnudde and Boone 2013] showed in their review work how much computed microtomography techniques have become indispensable in the field of geosciences due to great technological advances and computational progress. Macdonald [Macdonald et al. 2022], for example, used MicroCT to measure and characterize lithic tool and mathematically calculate edge curvature.

From the use of images for the analysis of fracture characteristics, Xiong et al. [Xiong et al. 2021] applied image processing techniques to investigate, for example, preferential fluid migration paths through the extraction of the three-dimensional fracture network present in the samples. In order to analyze how fractures propagate and interact, Li et al. [Li et al. 2022] used a three-dimensional reconstruction. Several procedures, such as stacking, recording and segmentation of the successive images of serial cuts, were also applied.

When image processing is used to extract information from the data, one of the most important steps is segmentation. To characterize the fracture geometry, Deng et al. [Deng et al. 2016] created their own segmentation method based on an interactive local

thresholding technique. This method significantly increased the number of classes in the gray scale and was able to isolate the fractures of the rock grains, thus making it possible to study the permeability caused by the fractures.

In the field of computational tools, Hardebol and Bertotti [Hardebol and Bertotti 2013] developed a software in Python programming language, called DigiFract, which works through digital photographs acquired from fracture outcrops, offering support for field acquisition, based on a core of the geographic information system (GIS). Healy et al. [Healy et al. 2017] created a tool called FracPaQ, implemented in MATLAB, which also aims to quantify the pattern of fractures in digital images, such as micrographs, geological maps, outcrop photographs, aerial or satellite images. On the other hand, Ramandi et al. [Ramandi et al. 2022] developed a novel algorithm called FracDetect that applies an automatic detection of fractures from grayscale 3D micro-CT images.

This type of work has demonstrated the importance of tools that allow the analysis and extraction of attributes such as the mechanical properties (for instance, strength and anisotropy) and transport ones (for instance, fluids and heat).

The understanding of these works indicates the great importance in the use of non-destructive methods for image analysis, helping to identify the most significant characteristics to extract the properties from the rock samples, making it possible to understand more clearly the needs of the study area. Therefore, we developed a methodology and a tool that work in a single environment with the objective of treating and analyzing the computed microtomographic images.

2.2. Related Concepts

The National Research Council [Council 1996] explains that the fundamental step in understanding and predicting fracture behavior involves identifying and locating hydraulically significant fractures. This is because the determination of its geometry is of paramount importance to feed flow simulations, which use mathematical models of flow in porous media to describe the physical phenomenon of mass and/or energy flow, based on macroscopic parameters [Ertekin et al. 2001].

Fractures have a series of individual characteristics that can be analyzed. Thus, to classify a sample that contains this structure, we raised the following indicators:

- The number of fractures identified in the image.
- The size of the fractures, which is a measure of the extent of development of the discontinuity surface. This carries the notion of size and controls the degree of fracturing.
- The type of connectivity the fracture has, which is a characteristic of the intersection and termination of a fracture. According to Ortega and Marrett [Ortega and Marrett 2000], connectivity is a fundamental property in terms of fluid flow. The quantification of this parameter would allow the comparison of different fracture networks and modeling of fluid flow through fracture systems in more realistic ways than commonly used. It can be classified into three types, as illustrated in Figure 2, taking into account the number of fracture connections as follows:

- Type I: not connected.
- Type II: individually connected.
- Type III: multiple connected.

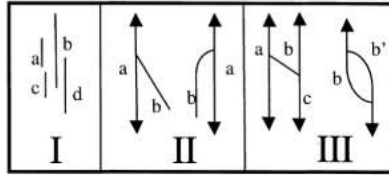


Figure 2. Fracture termination types. Isolated fractures (Type I, e.g., *b* fracture), individually connected fractures (Type II, *b* fracture), and fractures connected with other fractures in more than one location (Type III, *b* and *b'* fractures). Source: [Ortega and Marrett 2000].

Singhal and Gupta [Singhal and Gupta 2010] also highlighted that fractures can be identified in two types: systematic, which are planar and more regular in distribution, and non-systematic, which are irregular and curved, as illustrated in Figure 3.

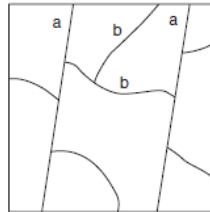


Figure 3. Representation of fracture types. (a) systematic and (b) non-systematic. Source: [Singhal and Gupta 2010].

- The fracture orientation, which can be classified into vertical, horizontal and inclined. This feature generally restricts the potential directions in which fluids can flow in a fractured rock system [ITRC 2020].
- The fracture density that corresponds to the degree/intensity of fracturing of the rock, which can be described in three ways: linear, regional and volumetric.
 - Linear fracture density is the average number of fractures in a given set, measured in a direction perpendicular to the fracture plane.
 - Regional fracture density (2D fracture density) is a way to quantify the persistence of discontinuity. It refers to the average length of strokes per unit area on a flat surface.
 - Volumetric fracture density (3D fracture density) is the average fractured surface area per unit rock volume created by all fractures in a given set.

3. Methodology

In this section, the methodology developed for the execution of fracture methods is described in detail. The developed methods were applied to public databases available on the Internet.

3.1. Segmentation

The segmentation of the images is performed in two steps until reaching the final result. The first step is to create a mask that indicates the region of interest of the image, while the second step produces an image referring to the rock content. Thus, by combining these results, an image can be created with only the internal structures that will be used for the next analyses.

To create the mask, a series of image processing techniques were applied using functions from the `Scikit-Image` [van der Walt et al. 2014] library and the `OpenCV` [Itseez 2015].

In Figure 4, it is possible to visualize a diagram that exemplifies the process of applying these algorithms to perform the segmentation. In it, the segmentation starts with the application of Otsu's thresholding in the original image, to extract the best threshold that separates the intensities of the image. Then, a labeled image is created with the object/background regions separated from the threshold obtained, using the following criteria:

$$label(x) = \begin{cases} 1, & \text{if } original(x) < \text{Otsu's thresholding} \\ 2, & \text{if } original(x) \geq \text{Otsu's thresholding} \end{cases} \quad (1)$$

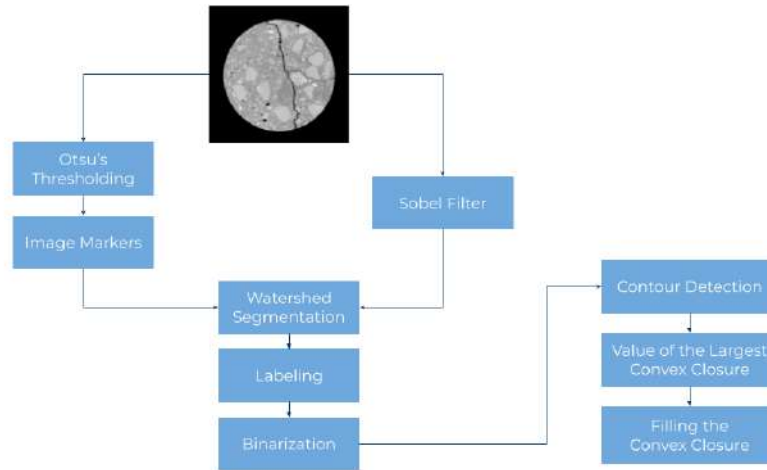


Figure 4. Diagram that illustrates the segmentation process.

It is also necessary to create an elevation map, obtained through the Sobel filter, to be used as a segmentation parameter using the watershed technique along with the labeled image. The creation of this map is a critical step for a good segmentation and helps to define the barriers of image structures.

After this procedure, the labeling technique is applied to the resulting segmented image in order to mark its regions and generate a new binary model, such that

$$binary(x) = \begin{cases} 0, & \text{if } labeling(x) = 0 \\ 255, & \text{if } labeling(x) \neq 0 \end{cases} \quad (2)$$

From this result, in which the region of interest of the image is highlighted, the contours are detected and an analysis of their convex closures is carried out to extract the one with the largest size, referring to the polygon provided by the area of the rock. Thus, the mask image is formed by filling the largest convex clasp, generating an output with the area relative to the rock in white and its background in black.

This detailed segmentation process is necessary because the rock sample usually has pores/fractures on its edge, such that a simple binarization would make it impossible to count and extract the geometry of these structures, since they would not have an edge that delimited them, being this are confused with the background, as shown in the third image in the sequence shown in Figure 5. Thus, with the creation of this mask in a circular format encompassing the entire sample area, all structures located in this region of interest are counted.

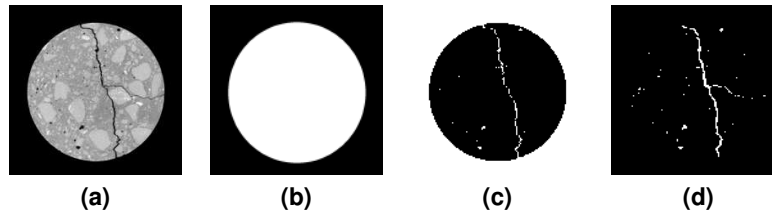


Figure 5. Microtomographic image and the resulting segmentation. (a) original image, (b) mask image, (c) image generated by the Otsu's thresholding and (d) image resulting from the two previous ones.

After creating the mask, another image is created by segmenting the original image by the Otsu's thresholding, and both images, when used together, provide greater confidence in capturing fractures. An example of these two techniques can be seen in Figure 5. In the end, a last image is created based on the previous two to provide a view of the pores/fractures only. Thus, where there is a white pixel in both images, the mask and the threshold, the white color is also inserted in the third image; otherwise, the color black is entered.

3.2. Feature Extraction

After segmenting the image, it is necessary to find the contours that were extracted from the original image in order to analyze its characteristics. For this, we use the *findcountours* function of the *OpenCV* library, where we pass the segmented image as a parameter and it returns the contours that were identified.

Fractures have an elongated body similar to a straight line, so we use this criterion to distinguish between other structures such as pores. One way to identify this feature is by applying the Hough line transform to a subimage extracted by the bounding box from the contour body. If a line can be drawn, this contour is classified as a fracture, otherwise it is excluded from the next steps. In Figure 6, an example of the operation of this process can be observed, applied to fractures identified in a coal sample.

From this filtering, the characteristics described in Subsection 2.2 are analyzed. The Hough line transform, in addition to allowing the identification of fractures, also makes it possible to calculate their slope through the equation of the distance between

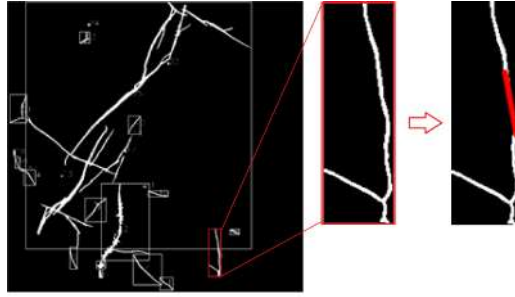


Figure 6. Line detection by the Hough line transform.

two points $(x1, y1)$ and $(x2, y2)$ in degrees, expressed as

$$angle = \arctan\left(\frac{y1 - y2}{x1 - x2}\right) * \frac{180}{\pi} \quad (3)$$

Characteristics, such as area, height and width, can be obtained through the same function that extracts the contour, as its return contains some data about the object's geometry. The calculation of the densities was based on the concepts defined in Section 2 and it is obtained through the pixels that compose the fracture. For linear density, the number of fractures by the diameter of the rock sample is used, given by

$$d1 = \frac{\text{number of fractures}}{\text{diameter}} \quad (4)$$

In 2D density, we calculate it by averaging the width of the fractures by the total area of the sample, that is

$$d2 = \frac{\text{mean}(\text{width of fractures})}{\text{total sample area}} \quad (5)$$

Finally, the 3D density, which calculates the ratio of the fracture area to the sample area, is expressed as

$$d3 = \frac{\text{total area of fractures}}{\text{total sample area}} \quad (6)$$

For the characterization of the fracture structure, for example, its connectivity and its type, the analysis must be more sophisticated due to its particularities, mainly due to the fact that many fractures can be close and interfere with an individual analysis. Therefore, two methodologies were developed to define these criteria:

- **Connectivity:** the first step to identify fracture bifurcations was to traverse each column of the non-rotating bounding box of each fracture in particular. Thus, in this process, it would be possible to count how many transitions occur from intensity relative to the fracture (white) to the background of the image (black). If there were more than one transition on the same line, this would indicate that there would be a bifurcation, raising its classification type (Types I, II, III) with each transition found. However, as shown in Figure 7, there are a number of

fractures that have parts of others inserted in their bounding box, which brings an incorrect result to the developed algorithm. Based on the study of some methods for identifying connected components (CC), the labeling technique was applied, in which, from this concept, it assigns the same label to a CC, making it possible to apply the initial algorithm again, but this time with the structures of a labeled bounding box. Thus, the fracture with the largest domain receives a label equal to 1 and the background of the image receives a label 0, allowing transitions to be counted only when they occur between these two values. Thus, if there is another fracture in the bounding box, it will receive another label, with the number 2 (shown in blue), not interfering with the result.

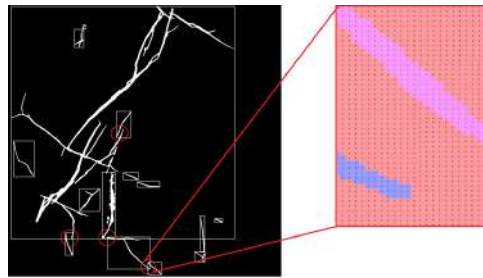


Figure 7. Identification of external components in the bounding box of a fracture.

- Type: To classify the fractures as systematic or not, the thickness of several fracture styles was studied. Through this, it was possible to perceive that those that were systematic, that is, did not have bifurcations, had a width smaller than 60 pixels. So it was used this value as a threshold for this parameter.

3.3. 3D Visualization

To obtain a complete view of the segmented images and to understand the distribution of the structures of interest, we implemented a function that allows the 3D visualization of these samples from the VTK tool.

VTK allows the reading of images in TIFF format through the `vtkTIFFReader()` class. After reading the images, we define a transfer function to map a property (in these cases, a scalar value) to a color value *Red-Green-Blue* (RGB) by the `vtkColorTransferFunction()` class. We also defined the opacity feature that will allow only white structures to be visualized, while the black color of the image is discarded by adding transparency.

An effective volume mapping is then generated in 3D space by the `vtkSmartVolumeMapper()` class and added to the rendering screen. In order to provide an interactive experience for the user and the sample, we integrated to this screen an interactive object of the `QVTKRenderWindowInteractor()` class, which allows the user to have different views of the sample by the movements performed with the mouse when pressing the rendered object.

4. Results

As a result of the models constructed for fracture analysis in computed microtomography rock images, we implemented an interface that allows user interaction with the data

obtained through the implementation of the methods described in the previous section.

For a better understanding of the real scale of this data, a transfer of pixel-millimeter metrics can be done from user interactions with the interface, indicating the value to which the pixel corresponds; if this data is unknown, the value 1 can be entered and the result will be in pixels.

In the area developed for the fractures, shown in Figure 8, all processing data are visualized. This data can be changed through the SliceBar, where the user can choose in which slice he/she intends to view the information. Thus, after the user sends the block of slices corresponding to his/her sample, two images are displayed on the screen, one referring to the original image and the other referring to the segmentation of the first slice of the block.

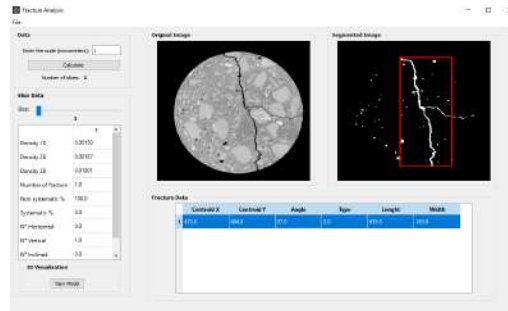


Figure 8. Interface screen to analyze and visualize fractures.

Along with the images, fracture data is also displayed in two tables. The vertical table shows the data corresponding to the slice as a whole, highlighting the following characteristics: the number of fractures in the slice, the densities of the slice (1D, 2D and 3D), the number of vertical, horizontal and inclined fractures, and the percentage of systematic and non-systematic fractures. In the horizontal table, located below the two images, it is possible to visualize the data of each fracture, indicated by the rows of the table. For a better correspondence of the numbers with the images, when clicking on the rows of the table, the user can see, in the segmented image, a mark on the fracture to which that data refers.

Finally, it is possible to obtain a 3D view of the input slices, where the user can interact by viewing their sample from different angles, as illustrated in Figure 9.



Figure 9. Different views obtained by 5 slices, passed as input, on the 3D view screen.

5. Conclusions

The characterization of reservoirs aims to describe the partial distribution of their petrophysical parameters, whose process is of paramount importance for the construction of a model that directs the development and management of an oil or gas field.

This paper presented a methodology to characterize the fracture geometry with computed microtomography rock images. In this sense, a computation tool was developed in which fractures can be identified and described, allowing an analysis of their behavior and shape, in order to contribute to a non-destructive evaluation of samples that represent a set under study.

Acknowledgments

The authors would like to thank CNPq (#309330/2018-1) and FAPESP (#2017/12646-3) for their support.

References

- Al Shafloot, T., Kim, T. W., and Kovsky, A. R. (2021). Investigating Fracture Propagation Characteristics in Shale Using sc-CO₂ and Water with the Aid of X-ray Computed Tomography. *Journal of Natural Gas Science and Engineering*, 92:103736.
- Basan, P. B., Lowden, B. D., Whittler, P. R., and Attard, J. J. (1997). Pore-Size Data in Petrophysics: A Perspective on the Measurement of Pore Geometry. *Geological Society, London, Special Publications*, 122(1):47–67.
- Cnudde, V. and Boone, M. N. (2013). High-resolution X-ray Computed Tomography in Geosciences: A Review of the Current Technology and Applications. *Earth-Science Reviews*, 123:1–17.
- Council, N. R. (1996). *Rock Fractures and Fluid Flow: Contemporary Understanding and Applications*. National Academies Press.
- Deng, H., Fitts, J. P., and Peters, C. A. (2016). Quantifying Fracture Geometry with X-ray Tomography: Technique of Iterative Local Thresholding (TILT) for 3D Image Segmentation. *Computational Geosciences*, 20(1):231–244.
- du Plasis, A. (2014). Concrete Cracking. <http://blogs.sun.ac.za/ctscanner/concrete-cracking/>.
- Ertekin, T., Abou-Kassem, J. H., and King, G. R. (2001). *Basic Applied Reservoir Simulation*, volume 7. Society of Petroleum Engineers Richardson.
- Hardebol, N. and Bertotti, G. (2013). DigiFract: A Software and Data Model Implementation for Flexible Acquisition and Processing of Fracture Data from Outcrops. *Computers & Geosciences*, 54:326–336.
- Healy, D., Rizzo, R. E., Cornwell, D. G., Farrell, N. J., Watkins, H., Timms, N. E., Gomez-Rivas, E., and Smith, M. (2017). FracPaQ: A MATLAB Toolbox for the Quantification of Fracture Patterns. *Journal of Structural Geology*, 95:1–16.
- ITRC (2020). Characterization and Remediation in Fractured Rocks. <https://fracturedrx-1.itrcweb.org/3-hydrology-fluid-flow/>.

- Itseez (2015). Open Source Computer Vision Library. <https://github.com/itseez/opencv>.
- Jing, L. and Stephansson, O. (2007). *Fundamentals of Discrete Element Methods for Rock Engineering: Theory and Applications*, volume 85. Elsevier.
- Li, M., Magsipoc, E., Abdelaziz, A., Ha, J., Peterson, K., and Grasselli, G. (2022). Mapping Fracture Complexity of Fractured Shale in Laboratory: Three-Dimensional Reconstruction from Serial-Section Images. *Rock Mechanics and Rock Engineering*, 55(5):2937–2948.
- Li, W., Frash, L. P., Welch, N. J., Carey, J. W., Meng, M., and Wigand, M. (2021). Stress-dependent Fracture Permeability Measurements and Implications for Shale Gas Production. *Fuel*, 290:119984.
- Lucia, F. J., Kerans, C., and Jennings, J. W. (2003). Carbonate Reservoir Characterization. *Journal of Petroleum Technology*, 55(06):70–72.
- Macdonald, D. A., Bartkowiak, T., Mendak, M., Stemp, W. J., Key, A., de la Torre, I., and Wieczorowski, M. (2022). Revisiting Lithic Edge Characterization with MicroCT: Multiscale Study of Edge Curvature, Re-Entrant Features, and Profile Geometry on Olduvai Gorge Quartzite Flakes. *Archaeological and Anthropological Sciences*, 14(2):1–20.
- Ortega, O. and Marrett, R. (2000). Prediction of Macrofracture Properties using Microfracture Information, Mesaverde Group Sandstones, San Juan Basin, New Mexico. *Journal of Structural Geology*, 22(5):571–588.
- Ramandi, H. L., Irtza, S., Sirojan, T., Naman, A., Mathew, R., Sethu, V., and Roshan, H. (2022). FracDetect: A Novel Algorithm for 3D Fracture Detection in Digital Fractured Rocks. *Journal of Hydrology*, 607:127482.
- Santos, A., Scanavini, H. F., Pedrini, H., Schiozer, D. J., Munerato, F. P., and Barreto, C. E. (2022). An Artificial Intelligence Method for Improving Upscaling in Complex Reservoirs. *Journal of Petroleum Science and Engineering*, 211:110071.
- Singhal, B. and Gupta, R. (2010). *Applied Hydrogeology of Fractured Rocks*. Springer Science & Business Media.
- van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. (2014). SciKit-Image: Image Processing in Python. *PeerJ*, 2:e453.
- Xiong, F., Jiang, Q., and Xu, C. (2021). Fast Equivalent Micro-Scale Pipe Network Representation of Rock Fractures Obtained by Computed Tomography for Fluid Flow Simulations. *Rock Mechanics and Rock Engineering*, 54(2):937–953.
- Yang, B., Wang, H., Wang, B., Shen, Z., Zheng, Y., Jia, Z., and Yan, W. (2021). Digital Quantification of Fracture in Full-Scale Rock Using Micro-CT Images: A Fracturing Experiment With N₂ and CO₂. *Journal of Petroleum Science and Engineering*, 196:107682.
- Zhu, W., Khirevich, S., and Patzek, T. W. (2021). Impact of Fracture Geometry and Topology on the Connectivity and Flow Properties of Stochastic Fracture Networks. *Water Resources Research*, 57(7):e2020WR028652.

Fire Patterns in the Pantanal Biome

Thais Pereira de Medeiros¹, Ana Larissa Ribeiro de Freitas², Maria Lucia Ferreira Barbosa¹, João Bosco Coura dos Reis², Galia Selaya³, Liana Oighenstein Anderson²

¹ Division of Earth Observation and Geoinformatics, National Institute for Space Research (INPE), São José dos Campos, SP.

² National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN), São José dos Campos, SP.

³ Ecosconsult-Prodigy, Santa Cruz, Bolivia.

{thais.pereira@inpe.br, alarisig@gmail.com, maria.lucia@inpe.br, joao.reis@cemaden.gov.br, gselaya@outlook.com, liana.anderson@cemaden.gov.br}

Abstract.

The Pantanal biome, the largest continental wetland in the world, has been experiencing an increase in fire occurrence, culminating with 2020 widespread wildfires. The main goal of this work was to answer the following question: What burns and how are the fire patterns in the Pantanal biome characterized? Our results indicated that the year with the highest burned area was 2016, with 13069.68 km² fire-affected areas. The peak month occurred in September, similar to the entire time series studies. Forest and pastures were the most affected, with 7319.52 km² and 2733.96 km², respectively. Understanding these patterns is essential to the development of public policies and the preservation of biomes and ecosystems.

1. Introduction

Fire events have become one of the major global concerns in recent years, given the increased recurrence of these events and their contribution to CO₂ emissions in the atmosphere, contributing significantly to global climate change (Barlow et al., 2012). Every year, mainly during the dry season, thousands of hectares of biomes and ecosystems are burning and, consequently, subject to irreparable damage. The fire occurrence and intensity are influenced by some factors, for instance, land use and land cover changes, and climatic conditions. Among the factors that influence these events, we can cite the elements that compose the fire triangle: [1] weather conditions, which includes atmospheric stability, wind (controls the speed and direction of fires), air temperature, precipitation and relative humidity; [2] fuel and flammability, which are influenced by chemical composition, structure and arrangement, spatial continuity, density of fuel and type of vegetation; [3] topography, which includes slope, aspect and elevation, and controls the speed of fires (Pivello et al., 2021; Burton, et al., 2020; UNEP, 2022).

Historically, fire has been used as a tool for forest clearing to convert it to agriculture and pasture. Fire can pose a high potential for forest degradation and ecosystem destruction, and a danger to populations and climate (Anderson et al., 2020).

The disturbance caused by these events can result in a large range of destructive ecological and social negative socioecological impacts, causing tree mortality; changes in forest structure and composition; long-term reduction in carbon stocks; economic losses; harm to the health of the local populations; impact on air quality, clouds formation, precipitation, hydrological cycle, and moisture (Chuvieco et al., 2013). Also, these consequences of wildfires can in the long-term reverse the progress of the United Nations Sustainable Development Goals (SDGs) and cause damage to biodiversity conservation (Martin, 2019). In addition, an intense succession of wildfires can alter the infiltration and evapotranspiration rates, increasing, consequently, the susceptibility of the area to the erosion processes.

There are many uncertainties about how and when the changes related to the occurrence and intensification of fires will occur. According to Jolly et al. (2015) under 1°C of global warming, the fire season has increased by an average of 33 days over the last 33 years in South America. According to a study developed by Burton et al. (2021), burned area is projected to increase in the future in all Shared Socioeconomic Pathways (SSP) scenarios, driven by an increase in temperature and a decrease in moisture. Understanding the factors that can alter the fire regime of an ecosystem is important and can help determine potential changes in fire behavior and, thus, their impacts of fire upon an ecosystem over a given period. The term fire regime includes the understanding of the role of fire in the structure and function of ecosystems and are described in terms of fire characteristics, e.g., fire extent, season, frequency, intensity and severity (Cochrane, 2009).

In this context, the Pantanal biome, the largest continental wetland in the world, which covers the countries of Brazil (80%), Bolivia (19%), and Paraguay (1%), was highlighted in the media for presenting a high number of fire events in the last years. The fire dynamics in the Pantanal are extremely complex and heterogeneous, mainly in the distinct portions of this biome (Brazil, Paraguay, and Bolivia). In 2020, for example, according to a study developed by Marques et al. (2021) was observed the largest quantity of burned area in the Brazilian Pantanal, about 50% higher than 2019. This shows changes in the predominant fire regime and a high-frequency of uncontrolled and unprecedented fire events. Moreover, Barbosa et al. (2022) found that climate variables in 2020 had an atypical behavior (hotter and drier than the historical average) and that burned areas in 2020 were 200% greater than the average. This result emphasizes that the increase in fire events was not only a result of extreme weather but also a consequence of human activities (Barbosa et al. 2022).

Pantanal has fire-prone vegetation composed of grasses and trees, which is characterized by a flood regime playing a vital role in structuring vegetation and is considered a World Natural Heritage Site and a Biosphere Reserve (UNESCO) (Abreu et al., 2022). However, few published studies still approached the patterns of fire in the biome (Marques et al., 2021), mainly in the portions of Bolivia and Paraguay. Furthermore, remote sensing techniques represent a particularly useful tool for mapping and quantifying the patterns, sizes, and assessing the impacts of fires on a range of natural and social systems. Thus, the main goal of this work was to answer the following question: What burns and how is characterized the patterns of fires in the Pantanal biome?

2. Materials and methods

2.1. Study Area

The study focuses on the Pantanal biome (Figure 1), one of the most biodiverse seasonal flood savannas in the world and the largest remaining wetland area of natural vegetation. This biome is located in the Upper Paraguay River Basin and includes two types of environments: lowlands (Pantanal biome) and highlands (Plateau). The Plateau is located outside the biome boundaries, but can influence the dynamics of the water cycle. The rainfall distribution is characterized by two well-defined seasons: rainy (October to March) and dry (April to September). This distribution controls the flooding cycles of the Pantanal. According to Schulz et al. (2019) these cycles are being changed, mainly due to deforestation for agriculture/pasture, the construction of waterways, which contribute to the intensification of drier conditions favorable to fire spread.

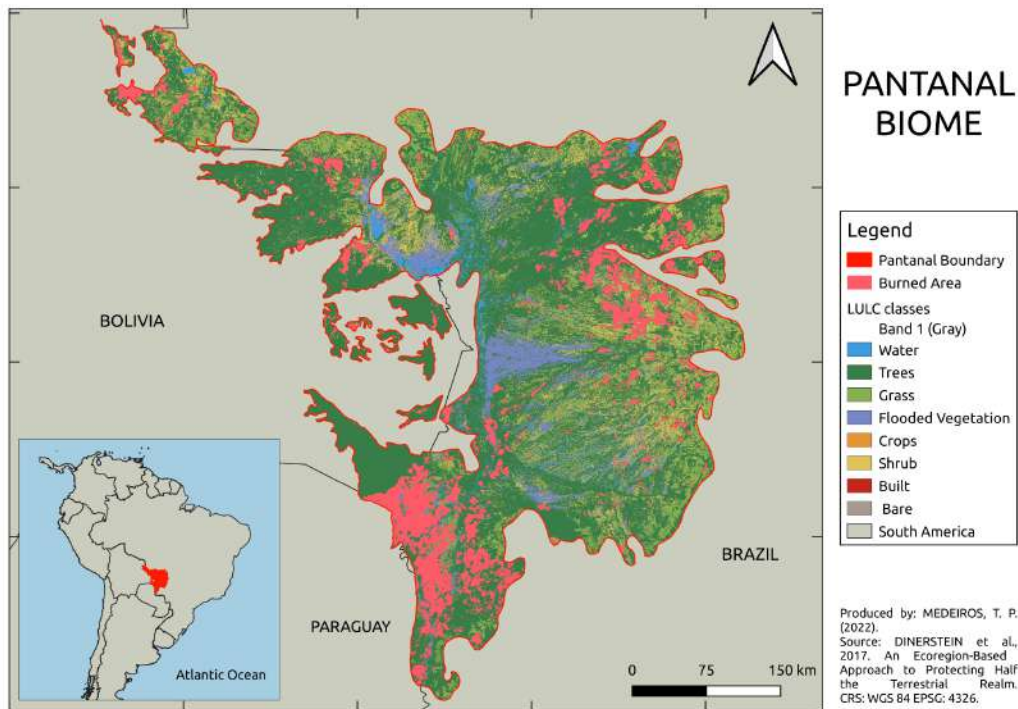


Figure 1. Study Area. Source: Dinerstein et al., 2017.

2.2. Data Collection

Aiming to understand the patterns of fires in the Pantanal biome, we collected the burned area dataset from the Global Fire Atlas (Andela et al., 2019). The Global Fire Atlas is a freely available global dataset that tracks the daily dynamics of individual fires to determine the timing and location of ignitions, fire size (km²), perimeter (km), duration (days), expansion (km²/days), fire line (fire line length, km), speed (km/day),

and direction of spread (dominant direction of spread, north, northeast, east, southeast, south, southwest, west, and northwest). In addition, aiming to understand what burned, we selected the Dynamic World Land Use and Land Cover dataset, a near real-time global 10 m land use and land cover mapping. This LULC classification was processed using Sentinel-2 imagery and deep learning techniques (Brown et al., 2022).

2.3. Methodological Procedures

The work was carried out in three main steps (Figure 2). The first step [1] was the extraction of metrics arising from Global Fire Atlas for the three years of the available time series (2016, 2017, and 2018): total burned area (fire size, km²) on the annual and monthly scales; monthly percentage of fire, expressed as the percentage ratio between the fire size in each month and the total fire size accumulated in the respective year. This calculation was also performed on the perimeter of the fire and fire line, both measured in km. In addition, we calculated the mean of duration (days), speed (km/day), and expansion (km²/day) of fire per month, as well as, the standard deviation of these metrics (duration, speed, and expansion). The second step [2] focused on quantifying what burned in each year, by intersecting the individual burn scars with the land cover classes maps.

Finally, the third step performs data analysis, we applied the LOESS Regression (Locally Weighted Scatterplot Smoothing), a non-parametric approach created to adjust a smooth curve to the data, where the noisy data values, sparse data points, or weak interrelationships influence the ability of creating a line with better adjustment. All graphics generated in this work were developed within a Python environment, using the PLOTLY library (<https://plotly.com/python/>), which making possible the creation of interactive graphs.

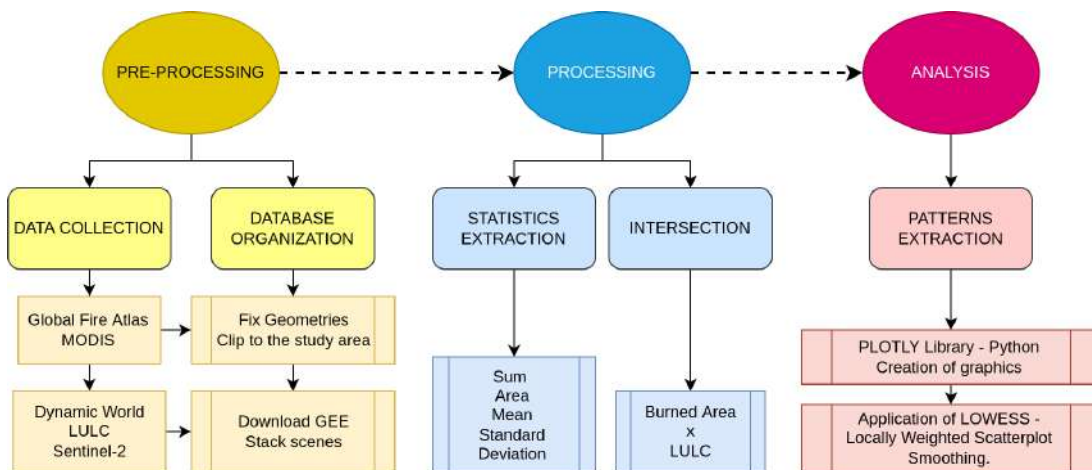


Figure 2. Flowchart of the methodological procedures developed in this work.

3. Results

The results are presented in two sections: [1] Monthly patterns of fires in the Pantanal biome according to metrics extracted from the Global Fire Atlas; [2] The identification of the most affected land use and land cover types.

The month and year with the largest burned area are September 2016 (~**6285 km²** burned), followed by September 2017 (~**5347 km²** burned) and October 2016 (~**3523 km²** burned). In addition, in 2016 we observed that September was responsible for **48%** of fire occurrences, followed by October, responsible for **27%** of fire occurrences, and finally, August, responsible for **14%**. In 2017, the month with the highest concentration of burn scars also occurred in September (**52%**), followed by August (**23%**), and July (**10%**). Finally, in 2018, the fire season is defined by two months: August (**48%**) and September (**44%**). The same fire pattern is visualized in graphics of Perimeter and Fire Line metrics (Figure 3). Therefore, 2016 was the year that burned the most and the fire season is mainly concentrated in July, August, September, and October.

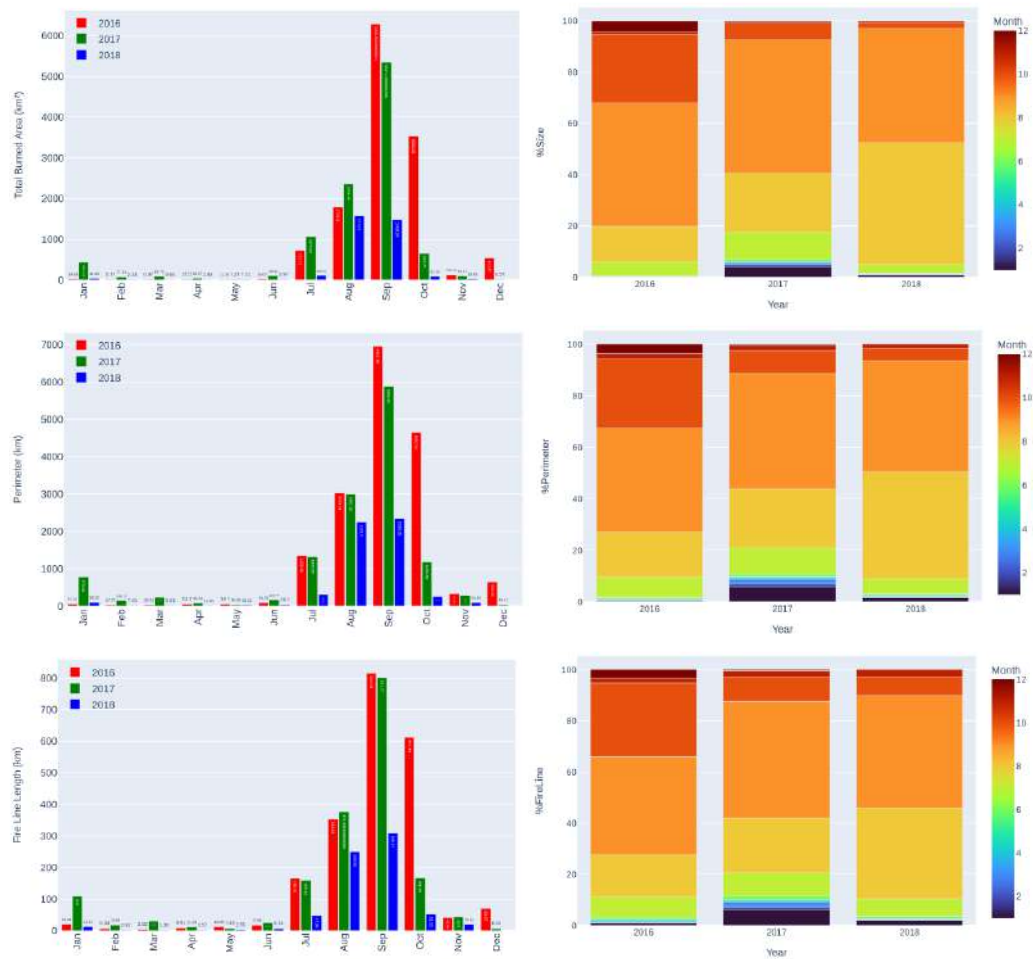


Figure 3. Total burned area per month and year - km² (upper left); Proportion of Burned Area per month - % (upper right); Total perimeter of the burn scars per

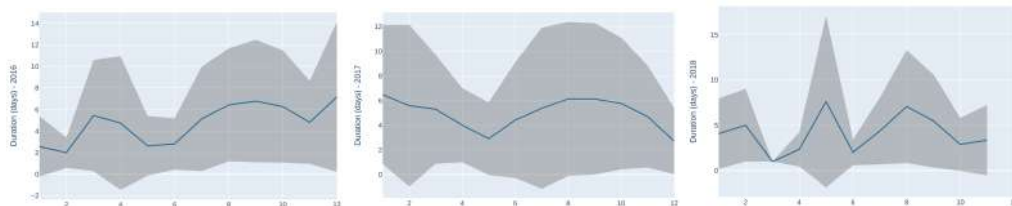
month and year - km (middle left); Proportion of Perimeter of the burn scars per month - % (middle right); Total length of the fire line per month and year - km (lower left); Proportion of Fire Line per month - % (lower right).

Other three metrics analyzed in this work are related to the duration of burn scars (days), speed of fire (km/day), and expansion of fire (km²/day) (Figure 4). Through the analysis of LOESS method, we found a similar distribution in Expansion and Speed of fire, where the higher mean values are observed in September in all years and the higher variability also occurs in the September (Table 1). In Table 1 we observed that all mean values exceed the annual mean values and presents a high standard deviation, that is, a high variability.

Table 1: Expansion (km²/day) and Speed (km/day) of fires over three years of time series (2016, 2017, 2018).

Fire metrics	2016	2017	2018
<i>Mean Expansion</i>	1.33 km ² /day	1.29 km ² /day	0.93 km ² /day
<i>Std Expansion</i>	5.08 km ² /day	5.08 km ² /day	2.52 km ² /day
<i>Annual mean Expansion</i>	0.82 km ² /day	0.86 km ² /day	0.71 km ² /day
<i>Mean Speed</i>	1.27 km/day	1.1 km/day	1.04 km/day
<i>Std Speed</i>	2.44 km/day	2.01 km/day	1.9 km/day
<i>Annual mean Speed</i>	0.71 km ² /day	0.71 km ² /day	0.44 km ² /day

However, the metric Duration of fire is more heterogeneous over the times series analyzed. For example, in 2016, the month with the highest mean days with fire was December, where each burn scar lasted an average of 7 days, while in 2017, the month with the highest value occurred in January (average duration of 6 days) and, finally, in 2018, the highlighted month was May (with an average duration of 8 days). Through this analysis, we can consider the month of September as the peak of fire in these years (2016, 2017, 2018) in the Pantanal biome, because the fire size, perimeter, fire line, speed, and expansion of fire, reached their maximum in this month.



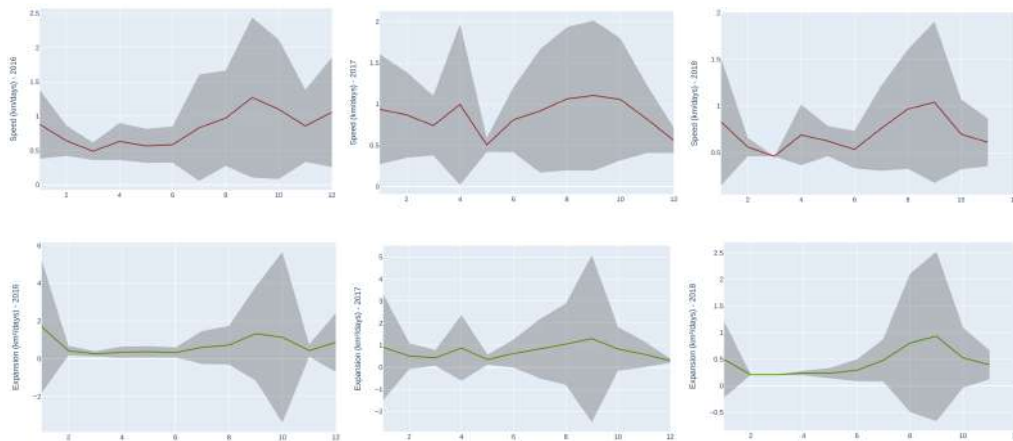


Figure 4. Average fire duration (days) over the months for the three years - 2016, 2017, 2018 (first row); Average fire speed (km/day) over the months for the three years - 2016, 2017, 2018 (second row); Average fire expansion (km²/day) over the months for the three years - 2016, 2017, 2018 (third row).

Analyzing the intersection performed between burned area and LULC data, we observed that during 2016, the land cover type most affected were Trees (37%), Grass (31%) and Shrub (30%), in 2017 was Trees (56%), Shrub (35%) and Grass (8%), and finally, in 2018 was Trees (68%), Shrub (27%) and Grass (3%). It is observed an increase in the Trees burned over the years and a decrease in the Grass burned (Figure 5).

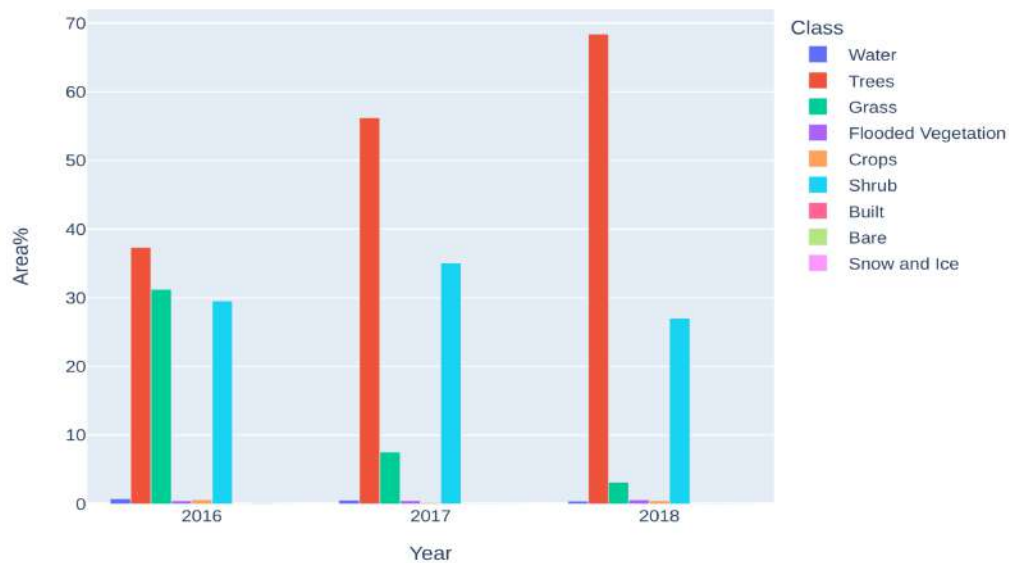


Figure 5. Percentage of burned area per Land Use and Land Cover (LULC) classes for the three years - 2016, 2017, 2018.

4. Discussion

Like the Brazilian savanna (Cerrado), a large portion of the Pantanal region is fire-prone, while the non-savanna portion of Pantanal is fire-sensitive, a region with no adaptation to fire events and susceptible to a plethora of disturbances (Hardesty et al., 2005). The fire pattern is defined in terms of size, frequency, or duration of burn scars, and their temporal variability is explained by rainfall seasonality, interannual variability and long-term droughts. Therefore, variations in precipitation rates along the seasons and the presence of droughts can affect the behavior of fires (Argibay et al., 2020).

The fire season in the Pantanal biome coincides with the dry period (April to September), extending until October (the first month of the wet season). Historically, fires in the Pantanal have been used as a low cost tool to convert natural vegetation into pasture and agriculture, bringing on exotic grasses, which can contribute to the alteration in the fire regime, this fact corroborates with one of the main causes of fires in the Pantanal, anthropogenic causes. Associated with anthropogenic causes, prolonged dry seasons and the intensification of extreme weather events, e.g., the 2016 drought, can influence the propagation of fire events (Barbosa et al., 2022).

In this context, it is possible to cite the ENSO (El Niño-Southern Oscillation) as a cause of 2016 drought, which is characterized by a Sea Surface Temperature (SST) anomaly, centered on the equatorial Pacific Ocean and divided between unusually warm SSTs (El Niño) or unusually cold SSTs (La Niña). These events can have a large impact on regular rainfall patterns and temperatures. The El Niño phase, for example, results in higher temperatures and reduced precipitation across the tropics, leading to increases in fire susceptibility. The increase in fire occurrence in the Pantanal during 2016 can be attributed to the peak of the last El Niño (Libonati et al., 2020; Burton et al., 2020).

According to Filho et al., (2021), many canals, called popularly “mouths”, have been blocked by farmers aiming to avoid the flooding of fields used for grazing, causing the difficulty of water spread and runoff, which, in turn, increases exposure to droughts. The natural climate conditions (dry and hotter conditions), associated with human activities and inadequate agriculture practices have been influencing the water supply of rivers, reducing their sizes and, consequently, increasing the susceptibility to fire events. In addition, the lack of consolidated legislation to protect the ecosystems and the lack of respect for the Protected Areas increases the natural and humans pressures within this biome, and contributing with their degradation.

4. Conclusions

The Pantanal biome is controlled by the alternation of consecutive years of flood and drought, which constitutes one of the most important actors in the socio-economic and biodiversity of the region. Fire events put such a balance at risk, by resulting in a large range of destructive ecological and social impacts. Through the analysis of fire metrics (size, perimeter, duration of fires, speed and expansion) we observed that the year with the highest burned area was 2016, coinciding with the year of the El Niño event, proving that most drier and hotter years can lead to a more flammable landscape. In addition, we saw that the fire peak (the month with the highest percentage of burned

area) occurred in September. These analyses were important because we could extract the main temporal patterns of fires, and identify what most burned, in this case, Forest (Trees) and Pasture (Grass). Identifying these patterns is essential for the development of public policies, aiming to the preservation of ecosystems and the reduction of impacts caused by fire events. Future works must be developed to understand the spatial patterns of fires and, thus, identify the areas most affected and susceptible to fire.

References

- Abreu, M. C. *et al.* (2022). Temporal and spatial patterns of fire activity in three biomes of Brazil. *Science of the Total Environment*, v. 884.
- Andela, N. *et al.* (2019). The Global Fire Atlas of individual fire size, duration, speed and direction. *Earth System Science Data*, v. 11, p. 529–552.
- Anderson, L. O. *et al.* (2021). Fire probability in South American Protected Areas Brazilian Settlements and Rural Properties in the Brazilian Amazon: December 2020 to February 2021. 32p. São José dos Campos, 2020. SEI/Cemaden processo 01250.029118/2018- 78/6265429.
- Argibay, D. S. Sparacino, J. Espindola, G. M. (2020). A long-term assessment of fire regimes in a Brazilian ecotone between seasonally dry tropical forests and savannah. *Ecological Indicators*, v. 113.
- Barbosa, M. L. F. *et al.* (2022). Compound impact of land use and extreme climate on the 2020 fire record of the Brazilian Pantanal. *Global Ecology and Biogeography*, p. 1-16.
- Barlow, J. *et al.* (2012). The critical importance of considering fire in REDD+ programs. *Biological Conservation*, v. 154, p. 1–8.
- Brown, C. F. *et al.* (2022). Dynamic World, near real-time global 10 m land use land cover mapping. *Scientific Data*, 9, 251.
- Burton, C. *et al.* (2020). El Niño Driven Changes in Global Fire 2015/16. *Frontiers in Earth Science*, v. 8, n. 199.
- Burton, C. Betts, R. A. Jones, C. D. Feldpausch, R. Cardoso, M. Anderson, L. O. (2020). El Niño Driven Changes in Global Fire 2015/16. *Frontiers in Earth Science*, v. 8, n. 199.
- Burton, C. Kelley, D. I. Jones, C. D. Betts, R. A. Cardoso, M. Anderson, L. (2021). South American fires and their impacts on ecosystems increase with continued emissions. *Climate Resilience Sustainability*.
- Chuvieco, E. Martínez, S. Román, M. V. Hantson, S. Pettinari, M. L. (2013). Integration of ecological and socio-economic factors to assess global vulnerability to wildfire. *Global Ecology and Biogeography*, v. 23, n. 2, p. 245–258.

- Cochrane. M. A. Fire in the tropics. In: Cochrane. M. A. (2009). *Tropical Fire Ecology: Climate Change, Land Use and Ecosystem Dynamics*. Springer Praxis Books. South Dakota State University, Brookings, USA.
- Dinerstein. E. et al. (2017). An Ecoregion-Based Approach to Protecting Half the Terrestrial Realm. *BioScience*, v. 67, n. 6, p. 534-545.
- Filho. W. L. Azeiteiro. U. M. Salvia. A. L. Fritzen. B. Libonati. R. (2021). Fire in Paradise: Why the Pantanal is burning. *Environmental Science and Policy*, v. 123, p. 31-34.
- Hardesty. J. Myers. R. Fulks. W. (2005). Fire, ecosystems, and people: a preliminary assessment of fire as a global conservation issue. *George Wright Forum*, v. 22, n. 4, p. 78-87.
- Jolly. M. et al. (2015). Climate-induced variations in global wildfire danger from 1979 to 2013. *Nature Communications*, v. 6, n. 7537.
- Libonati. R. et al. (2021). *Sistema ALARMES – Alerta de área queimada Pantanal, situação final de 2020*. Nota Técnica 01/2021. Rio de Janeiro: Laboratório de Aplicações de Satélites Ambientais.
- Marques. J. F. et al. (2021). Fires dynamics in the Pantanal: Impacts of anthropogenic activities and climate change. *Journal of Environmental Management*, v. 299.
- Martin. D. A. (2019). Linking fire and the United Nations Sustainable Development Goals. *Science of the Total Environment*, v. 662, p. 547-558.
- Pivello. V. R. et al. (2021). Understanding Brazil's catastrophic fires: Causes, consequences and policy needed to prevent future tragedies. *Perspectives in ecology and conservation*, v. 19, p. 233-255.
- Schulz. C. et al. (2019). Physical, ecological and human dimensions of environmental change in Brazil's Pantanal wetland: synthesis and research agenda. *Science of the Total Environment*, v. 687, 1011–1027.
- Sullivan. A. Baker. E. Kurvits. T. (Eds). *Spreading like Wildfire – The Rising Threat of Extraordinary Landscape Fires*. A UNEP (United Nations Environment Programme) Rapid Response Assessment. Nairobi, 2022.

Database of 2019 Brazilian oil disaster: an overview of a dataset and its application on an Artificial Intelligence monitoring system.

¹Maria Paula Graziotto; ²Laércio Massaru Namikawa

¹ -Bolsista DTI no Instituto Nacional de Pesquisas Espaciais - INPE

² -Tecnologista Sênior III no Instituto Nacional de Pesquisas Espaciais - INPE

{maria.graziotto@inpe.br; laercio.namikawa@inpe.br}

Abstract. In September 2019 small tar balls and oil slicks showed up affecting eleven states, mostly in the Northeast region, which depends mostly on tourism and fishery activities. The Brazilian National Contingency Plan (PNC) failed since the oil spread subsurface making impossible to see thorough satellite images. To help out to discover what happened, the Federal Government activated the Monitoring and Evaluation Group (GAA) to study different aspects of this disaster and learn how to prevent and mitigate a possible new event. This work is part of the Work Group 1, which deals with the Monitoring and Modeling, and describes the dataset composed of SAR images used to train the Artificial Intelligence (AI) and the future data that will be incorporated on the database of the project.

1 - Introduction

In September 2019 small tar balls and oil slicks showed up at over 1000 locations along the Brazilian's coastline till December of 2020 (IBAMA 2020). Altogether, it affected 11 states, mostly in the Northeast region (>80%) and 20% in the Southeast region (Soares et al. 2020). It represented the worst oil spill disaster not only in Brazilian history, but also in any tropical coastal region worldwide (Nasri Sissini et al. 2020).

The socioeconomic and environmental impacts of this event are not only a short term problem, but also a long term situation, making it difficult to estimate the price of the damage (Câmara et al. 2021). In terms of socioeconomic matters, those states depend on tourism and fishery activities. In the Northeast region approximately 724 fishery and mussels extraction locations were spoiled, affecting 144 thousand artisanal fishers (IBAMA 2020), jeopardizing their incoming for a long period. In less than 2 months of the disaster 40% of the tourism movement decayed (FUNDAJ 2019). Studies show the public health issues which those places are passing through, such as mental health, breathing problems and skin diseases (Carmo and Teixeira 2020); (Pena et al. 2020); (CCI/ENSP 2021).

Despite the Brazilian National Contingency Plan (PNC) existing, it didn't apply correctly in this case. Firstly because it should be triggered with an emergency, but it took weeks for the

government to act. Secondly, when specialists analyzed the satellite images to discover where and who caused the spill, they couldn't conclude what happened. This was because the oil spill spread subsurface, so any satellite or radar could not identify the exact location of the incident and inquire the indicted. With that scenario, the Federal Government activated the Monitoring and Evaluation Group (GAA), jointly coordinated by the Brazilian Institute of Environmental and Renewable Natural Resources (IBAMA), the Brazilian Navy and the National Petroleum Agency (ANP). To help out to discover where and who committed the spill, the GAA created a scientific coordination body composed of seven working groups (WGs) that collaborated with scientists working from all over the country to study different aspects of this disaster and learn how to prevent and mitigate a possible new event. To succeed in that, a combination of technologies will be used.

Remote Sensing (RS) has been developed in the past decade, becoming a powerful tool to monitor every spot in the planet with a high resolution and frequency of coverage. Multisensory data has been collected for many years, making the lack of data no longer a limit factor for Earth Observation (EO). The challenge nowadays is to analyze all that data. For that, Machine Learning and Artificial Intelligence came to promise maximization of the work and to make possible the monitoring of large areas 24/7.

This work is part of the WG 1, which deals with the Monitoring and Modeling, which aims to develop situational analysis and Artificial Intelligence algorithm for detection of oil slicks on the sea surface from orbital images, meteo-oceanographic fields of surface generated by coupled ocean-atmosphere models and ship positions. To do that, all data must be on an integrating platform that must be able to store, select and make these data available for analysis in a geographic information system. Analyzes must be carried out semi-automatically, using images processing algorithms associated with artificial intelligence techniques (Nobre et al. 2022). The aim of this paper is to describe the creation and dataset gathering for the Artificial Intelligence (AI) training and the future data that will be incorporated on the database of the project. This primary database set will nourish the ship and oil detection algorithms developed by WG that will monitor the Brazilian coast and help to prevent and mitigate any spill like the 2019 event.

2 - Assembling Data for the Artificial Intelligence

Oil spill is one of the biggest problems that affect the environment since the industrial revolution. Petroleum products have a wide range of short/ long-term impacts on the ecosystem. Animals exposed to oil can have health problems and changes to their physiology and behavior (Ober 2019). Fortunately, the amount of large spills declined over the decades. The International Tanker Owners Pollution Federation (ITOPF) maintains a database of oil spill from tank vessels since 1970, and shows the reduction of spills over the years. Despite that downward trend, due to improvement on the standards of operations in sea-borne transportation, most of those incidents occurred while the vessel was underway in open water (ITOPF 2022), where inspection is a problem.

Monitoring and detecting an oil slick as soon as possible is the key to a successful contingency plan. Satellite-based remote sensing became the monitoring into remote areas, as the middle of the ocean, possible. Radar images, like the Synthetic Aperture Radar (SAR) are the best instrument to do that since it provides data regardless of the hour of the day and the weather condition. Oil slick shows up as a dark area above a light gray color which represents water. However, dark areas in SAR images do not always mean petroleum slicks. These oil spill “look-alike” may represent: algae blooms; low wind track; cold upwelling water; divergent flow, e.g., wave or tidal flow; turbulent water; and other kinds of oil, e.g. plant oils (Alpers, Holt, and Zeng 2017).

One of the many techniques adopted to identify oil in SAR images is Machine Learn (ML). It has shown an effective way to extract information from RS data in a semi-automatic manner, turning the analyses of a massive amount of data faster, helping the right decision making on the time. Al-Ruzouq et al. (2020) made a great review of this technique showing the advantages and shortcomings, the preprocessing chain needed and the challenges of oil spill detection.

As previously said, to differ what is oil from oil’s looks-alike it is the big issue. To minimize that error, images from verified oily incidents were chosen within meteo-oceanographic data from those episodes.

For now, the dataset is composed of twenty-six RADARSAT-1 images from Cantarell (Gulf of Mexico) of 2020 (Figure 1). Cantarell is a heavy oilfield located 100km off the Yucatán coast, in Mexico. It was once the second largest oilfield in the world, but currently occupies the sixth position (Offshore Technology 2000; Wikipedia contributors 2022). All images were pre-processed with an image segmentation technique that analyzed 409 features of oil, classified in oil spill (349 features) or seepage slick (60 features) (Beisl et al. 2004; de Miranda et al. 2004).

Sixty Sentinel-1 images (Figure 2), being sixteen from the Suez Canal and four images from the Red Sea contained verified oil spill events, related to ship incidents, within the period 2015-2019 (Abou El-Magd et al. 2020). Five images from the oil spill accident with the Wakashio Japanese vessel in 2020 at Mauritius, an Indian Ocean island (Rajendran et al. 2021).

Four scenes from the period of 2017 to 2018, one close to Kamarajar Harbor, south India, where two ships collided. One at Sharjah, United Arab Emirates, due to an oil well drilling. Others from Kuwait coast, with an uncertain origin. And an image from Mubarak, Pakistan, related to an illegal discard (Naz et al. 2021). Two scenes from the Black Sea, where a tanker dropped oil, spreading for almost 80 km² on the sea in August 2021 (Voytenko 2021).

Thirteen images from Mediterranean Sea, seven from the Syrian oil refinery in Baniyas at August 2021, tree images of Emerald tanker oil spill that roamed from Israel to Lebanon coast in February 2021, and tree images from Cap Corse due to the collision between two vessels in 2018 (VisioTerra 2021; Surkes 2021; Bitelli 2019).

Eight scenes from the North Sea, near to the Brage oil platform, show water discharges containing oil in it (Skrunes, Johansson, and Brekke 2019). And four images from the Caspian

Sea, on the Azerbaijan coast. Two of them are from the fire incident on the gas pipeline in 2015, and two are from irregular discharges in 2019 (Mityagina and Lavrova 2019).

Images from other satellite sources like ICEYE, MODIS and CBERS, COSMO-SkyMed Constellation and TerraSAR-X with more oil spill cases will be incorporated, especially from the event of 2019. Also, wind, chlorophyll and sea surface data related to the images sites, from ERA5 and MODIS models are being added to the database along with real time data from PNBOIA/REMObs buoys to assist the AI to understand what is real oil spills from “look-alikes”. All images will be stored at GeoTiff and the meteo-oceanography data in NetCDF formats.

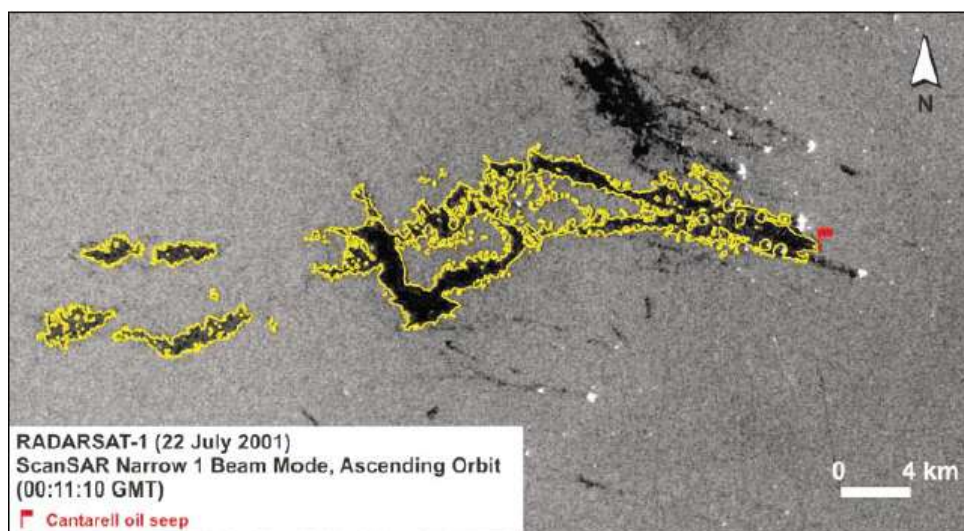


Figure 1. RADARSAT-1 image with seepage slick (highlighted in yellow) from Cantarell (2001).
Source: de Miranda et al. 2004.

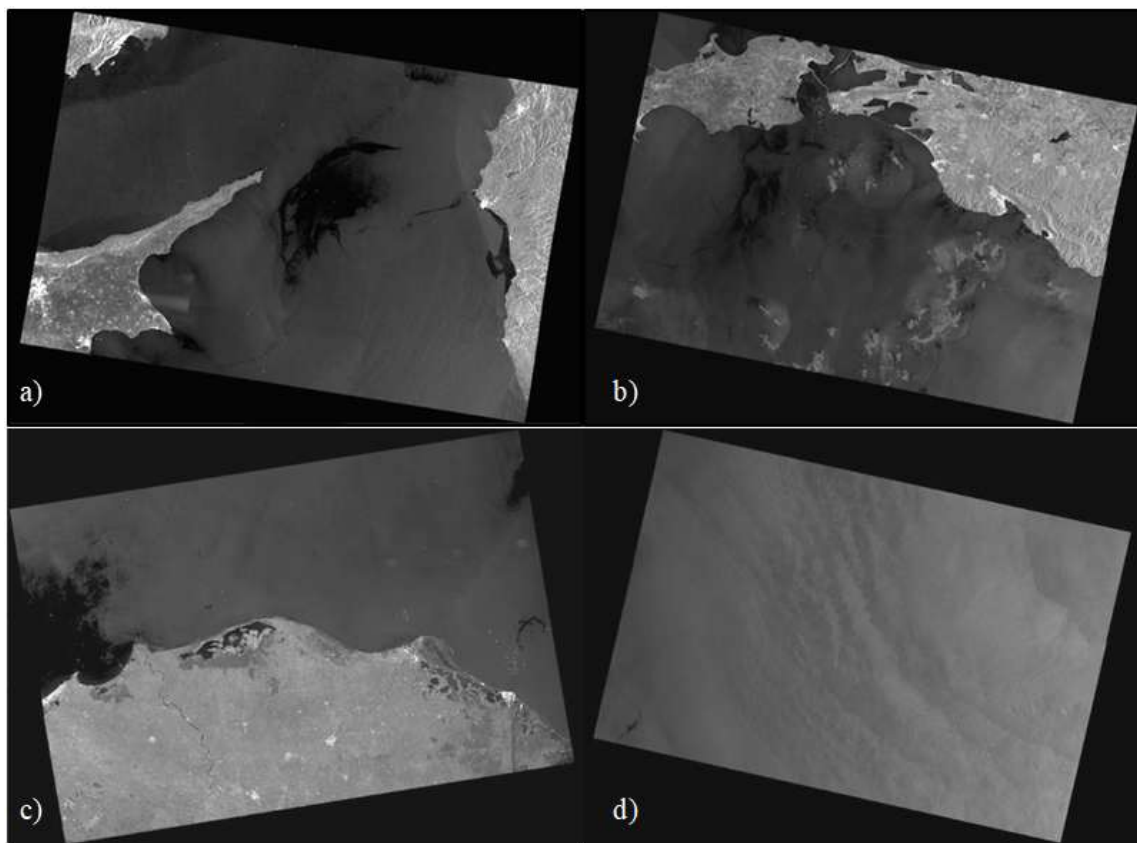


Figure 2. Sentinel-1 images with oil spills. a) Syrian coast 2021, b) Black Sea 2021 , c) Suez Canal 2019, d) North Sea 2019.

3 - Database Creation

To catalog the data, spreadsheets containing the necessary information to understand the environment of the image were made. One describes the metadata of the image and another describing the environment of the place at that time.

The information extracted from the images is:

-Location:

- ID_data - is the Mercator zone where the center of the image is. The name of the satellite and the date of acquisition. Used to connect with the others worksheets;
- ID_mercator: is the Mercator zone where the center of the image is.
- Lat_dec_central / Lon_dec_central - decimal degrees from the center of image;
- Sigla_uf - states there are in the image, if it is in Brazil territory.

-Images data:

- Satellite - the satellite name;
- Sensor - Instruments related to the acquisition of data or satellite mission. E.g.: MUX, WFI, SAR;
- Nome_arquivo_imagem - original name of the image;
- Bandas - refers to the spectral resolution at which the sensor operates. Usually they are: infrared, blue, green, red;
- Órbita - the orbit of a satellite depends on the function used for it. Some types of orbits are: circular elliptical, polar and geostationary;
- Data - the date of the acquisition;
- Process_Type - which level of processing it has undergoing;
- Status - if it has oil spill, seepages slicks or false oil;
- Features - how many oil features does the image contain.

The information extracted from the environment is:

-Locator:

- ID_data - worksheets connector;

-Input:

- Lat_dec / Lon_dec - decimal degrees from the center of site related to the image;
- Sigla_uf - if it is in Brazil territory, the states within;
- Data - date of the inputs data;
- Hora - hour of the data acquisition;
- Vento - minimum and maximum of wind on that quadrant;
- Mare - minimum and maximum of tide on that quadrant;
- Corrente - minimum and maximum of current speed on that quadrant;
- Batimetria - better source of bathymetry to run an oceanography model of the place;

To handle the capacity of storage and processing of all data that the project aims to deal with, a Spatial Data Infrastructure (SDI) was chosen. SDI is a platform that facilitates the management of EO data with an easy interface between the manager and the system. To improve

the efficiency and performance to process the large amount of data, those platforms are using their own system of Array Database Management System (Array DBMS) such as TileDB (Papadopoulos et al. 2016) and RasDaMan (Baumann et al. 1998). They also developed their own query language for multidimensional arrays, like AQL (Array Query Language) and RasDaMan RasQL. The better fit to this project was the Open Data Cube (ODC), described next.

3.1 Open Data Cube

To choose the better platform for the project, some requirements and characteristics need to be considered. Gomes et al. (2020) listed the capacities of the platforms, which helped to select it. For that project, the capability of storage scalability should be easy; other interfaces and applications must have access to data of other services, like OGC Services; the extensibility for new tools; the capacity of sharing and reproducing the results among others researchers; and the processing capability is essential.

The Open Data Cube (ODC) is an open source project born from the need to manage a bunch of Satellite Data. It supports through command line tools and a Python API the cataloging, access and manipulation of a huge EO datasets, distributed on git repositories (<https://github.com/opendatacube>). Figure 4 shows the whole process of uploading the data, the data cube infrastructure and application platform. To visualize products, statistics and retrieve datasets on GeoTiff and NetCDF formats, a web portal is available, which can be managed through a REST API on Python Jupyter Notebooks which makes easy to access and manipulate the data (Gomes, Queiroz and Ferreira 2020; Open Data Cube 2022).

At this moment, this platform is in the process of installation and configuration on the project supercomputer, and soon will be available for access.

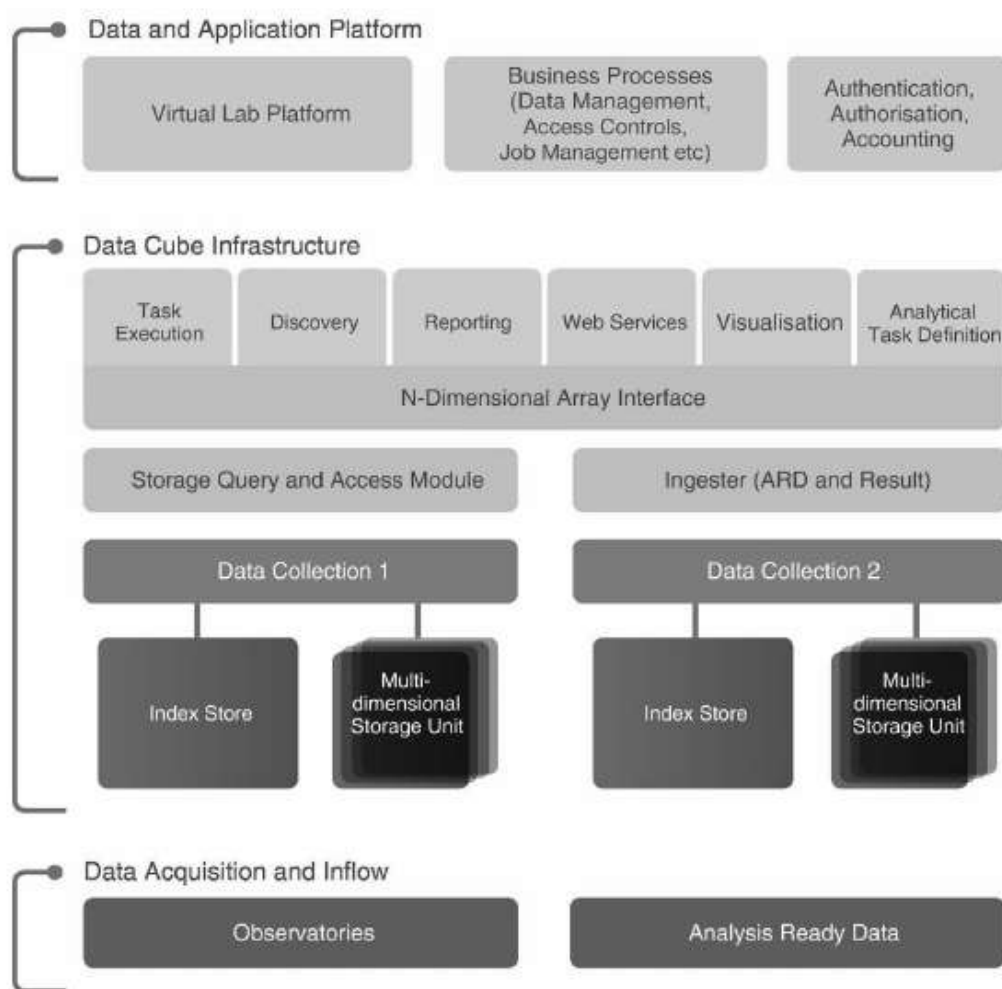


Figure 4. Open Data Cube (ODC) platform architecture diagram. Source: (Lewis et al. 2017)

4 - Consideration

This work belongs to a project that aims to create a lifetime monitoring system of Brazilian marine space using Artificial Intelligence and Satellite data. To do that many tests, trials and work will be necessary. The database's first steps are to collect information from the event of 2019 and other verified incidents to teach the AI to read oil in satellite images. Then, nourishing the DB with new oil spills data from our country and afford an integrated platform so users can have access to it, using the better SDI platform for that project.

5 - Acknowledges

The authors acknowledge the support of the National Institute for Space Research - INPE for support in the completion of this study. This work was supported by Conselho Nacional de

Desenvolvimento Científico e Tecnológico – CNPq Grant No. 440857/2020-1, CNPq/MCTI 06/2020 - Pesquisa e Desenvolvimento para Enfrentamento de Derramamento de Óleo na Costa Brasileira, Programa Ciência no Mar. The satellite data provided by Copernicus Marine Environment Monitoring Service - CMEMS and Open Data Cube are acknowledged. For our partner Carlos Beisl and his team to provide satellite data with segmentation of oil already done. And all the researchers and institutions who are part of this project.

6- References

- Abou El-Magd, Islam, Mohamed Zakzouk, Abdulaziz M. Abdulaziz, and Elham M. Ali. 2020. “The Potentiality of Operational Mapping of Oil Pollution in the Mediterranean Sea near the Entrance of the Suez Canal Using Sentinel-1 SAR Data.” *Remote Sensing* 12 (8): 1352. <https://doi.org/10.3390/rs12081352>.
- Alpers, Werner, Benjamin Holt, and Kan Zeng. 2017. “Oil Spill Detection by Imaging Radars: Challenges and Pitfalls.” *Remote Sensing of Environment* 201 (November): 133–47. <https://doi.org/10.1016/j.rse.2017.09.002>.
- Al-Ruzouq, Rami, Mohamed Barakat A. Gibril, Abdallah Shanableh, Abubakir Kais, Osman Hamed, Saeed Al-Mansoori, and Mohamad Ali Khalil. 2020. “Sensors, Features, and Machine Learning for Oil Spill Detection and Monitoring: A Review.” *Remote Sensing* 12 (20): 3338. <https://doi.org/10.3390/rs12203338>.
- Baumann, P., A. Dehmel, P. Furtado, R. Ritsch, and N. Widmann. 1998. “The Multidimensional Database System RasDaMan.” *SIGMOD Rec.* 27 (2): 575–77. <https://doi.org/10.1145/276305.276386>.
- Beisl, C. H., E. C. Pedroso, L. S. Soler, A. G. Evsukoff, F. P. Miranda, A. Mendoza, A. Vera, and J. M. Macedo. 2004. “Use of Genetic Algorithm to Identify the Source Point of Seepage Slick Clusters Interpreted from Radarsat-1 Images in the Gulf of Mexico.” *Institute of Electrical and Eletronics Engineers IEEE* 6: 4139–42. <https://doi.org/10.1109/IGARSS.2004.1370044>.
- Bitelli, Francesca Franci Francesca Trevisiol. 2019. “SENTINEL-1 DATA FOR FUEL SPILL DETECTION. The Case of the Ships Collision near Corsica.” https://www.enea.it/it/seguici/events/telerilevamento_25-26giu2019/Trevisiol.pdf.
- Câmara, Samuel Façanha, Francisco Roberto Pinto, Felipe Roberto da Silva, Marcelo de Oliveira Soares, and Thiago Matheus De Paula. 2021. “Socioeconomic Vulnerability of Communities on the Brazilian Coast to the Largest Oil Spill (2019–2020) in Tropical Oceans.” *Ocean & Coastal Management* 202 (March): 105506. <https://doi.org/10.1016/j.ocecoaman.2020.105506>.
- Carmo, Eduardo Hage, and Maria Gloria Teixeira. 2020. “Technological Disasters and Public Health Emergencies: The Case of Oil Spill on the Brazilian Coast.” *Cadernos de Saude Publica* 36 (2): e00234419. <https://doi.org/10.1590/0102-311X00234419>.
- CCI/ENSP. 2021. “Derramamento de óleo no litoral brasileiro: pesquisadores concluem que desastre configurou Emergência em Saúde Pública.” 2021. <http://informe.ensp.fiocruz.br/noticias/48246>.

- FUNDAJ. 2019. “5 fatos que não te contaram sobre a poluição por óleo no Litoral do Nordeste.” FUNDAJ. October 21, 2019. <https://www.gov.br/fundaj/pt-br/destaques/observa-fundaj-itens/observa-fundaj/revitalizacao-de-bacias/5-fatos-que-nao-te-contaram-sobre-a-poluicao-por-oleo-no-litoral-do-nordeste>.
- Gomes, Vitor C. F., Gilberto R. Queiroz, and Karine R. Ferreira. 2020. “An Overview of Platforms for Big Earth Observation Data Management and Analysis.” *Remote Sensing* 12 (8): 1253. <https://doi.org/10.3390/rs12081253>.
- IBAMA. 2020. “Manchas de Óleo - Litoral Brasileiro.” IBAMA. 2020. <http://www.ibama.gov.br/manchasdeoleo>.
- ITOPF. 2022. “Oil Tanker Spill Statistics 2021.” ITOPF. 2022. <https://www.itopf.org/knowledge-resources/data-statistics/statistics/>.
- Lewis, Adam, Simon Oliver, Leo Lymburner, Ben Evans, Lesley Wyborn, Norman Mueller, Gregory Raevksi, et al. 2017. “The Australian Geoscience Data Cube — Foundations and Lessons Learned.” *Remote Sensing of Environment* 202 (December): 276–92. <https://doi.org/10.1016/j.rse.2017.03.015>.
- Miranda, Fernando Pellon de, Arturo Mendoza Quintero Marmol, Enrico Campos Pedroso, Carlos Henrique Beisl, Pamela Welgan, and Luis Medrano Morales. 2004. “Analysis of RADARSAT-1 Data for Offshore Monitoring Activities in the Cantarell Complex, Gulf of Mexico, Using the Unsupervised Semivariogram Textural Classifier (USTC).” *Canadian Journal of Remote Sensing* 30 (3): 424–36. <https://doi.org/10.5589/m04-019>.
- Mityagina, Marina I., and Olga Y. Lavrova. 2019. “Multi-Sensor Satellite Survey of Surface Oil Pollution in the Caspian Sea.” In *Remote Sensing of the Ocean, Sea Ice, Coastal Waters, and Large Water Regions 2015*, edited by Charles R. Bostater, Stelios P. Mertikas, and Xavier Neyt. SPIE. <https://doi.org/10.1117/12.2194511>.
- Nasri Sissini, M., F. Berchez, J. Hall-Spencer, N. Ghilardi-Lopes, V. F. Carvalho, N. Schubert, G. Koerich, et al. 2020. “Brazil Oil Spill Response: Protect Rhodolith Beds.” *Science*. <https://doi.org/10.1126/science.aba2582>.
- Naz, Saima, Muhammad Farooq Iqbal, Irfan Mahmood, and Mona Allam. 2021. “Marine Oil Spill Detection Using Synthetic Aperture Radar over Indian Ocean.” *Marine Pollution Bulletin* 162 (January): 111921. <https://doi.org/10.1016/j.marpolbul.2020.111921>.
- Nobre, Paulo et al. The 2019 northeast Brazil oil spill: scenarios. *Anais da Academia Brasileira de Ciências* [online]. 2022, v. 94, suppl 2 [Accessed 13 September 2022] , e20210391. Available from: <<https://doi.org/10.1590/0001-376520220210391>>. Epub 02 Sept 2022. ISSN 1678-2690. <https://doi.org/10.1590/0001-376520220210391>.
- Ober, Holly K. 2019. “Effects of Oil Spills on Marine and Coastal Wildlife.” *Department of Wildlife Ecology and Conservation*. <https://edis.ifas.ufl.edu/publication/UW330>.
- Offshore Technology. 2000. “Cantarell Oilfield, Gulf of Mexico.” Offshore Technology. May 16, 2000. <https://www.offshore-technology.com/projects/cantarell/>.
- Open Data Cube. 2022. Available online: <https://www.opendatacube.org/overview> [Accessed: 20 June 2022].
- Papadopoulos, Stavros, Kushal Datta, Samuel Madden, and Timothy Mattson. 2016. “The TileDB Array Data Storage Manager.” *Proceedings of the VLDB Endowment International Conference on Very Large Data Bases* 10 (4): 349–60. <https://doi.org/10.14778/3025111.3025117>.
- Pena, Paulo Gilvane Lopes, Amanda Laura Northcross, Mônica Angelim Gomes de Lima, and Rita de Cássia Franco Rêgo. 2020. “The Crude Oil Spill on the Brazilian Coast in 2019: The

- Question of Public Health Emergency.” *Cadernos de Saude Publica* 36 (2): e00231019. <https://doi.org/10.1590/0102-311X00231019>.
- Rajendran, Sankaran, Ponnumony Vethamony, Fadhil N. Sadooni, Hamad Al-Saad Al-Kuwari, Jassim A. Al-Khayat, Vashist O. Seegobin, Himanshu Govil, and Sobhi Nasir. 2021. “Detection of Wakashio Oil Spill off Mauritius Using Sentinel-1 and 2 Data: Capability of Sensors, Image Transformation Methods and Mapping.” *Environmental Pollution* 274 (April): 116618. <https://doi.org/10.1016/j.envpol.2021.116618>.
- Schramm, Matthias, Edzer Pebesma, Milutin Milenković, Luca Foresta, Jeroen Dries, Alexander Jacob, Wolfgang Wagner, et al. 2021. “The openEO API—Harmonising the Use of Earth Observation Cloud Services Using Virtual Data Cube Functionalities.” *Remote Sensing* 13 (6): 1125. <https://doi.org/10.3390/rs13061125>.
- Skrunes, Stine, A. Malin Johansson, and Camilla Brekke. 2019. “Synthetic Aperture Radar Remote Sensing of Operational Platform Produced Water Releases.” *Remote Sensing* 11 (23): 2882. <https://doi.org/10.3390/rs11232882>.
- Soares, M.O, Teixeira, C. E. P., Bezerra, L. E. A., Paiva, S. V., Tavares, T. C. L., Garcia, T. M., Araújo, J. T., Campos, C. C., Ferreira, S. M. C., Matthews-Cascon, H., Frota, A., Mont’Alverne, T. C. F., Silva, S.T., Rabelo, E. F., Barroso, C. X., Freitas, J. E. P., Júnior, M. M., Campelo, R. P. S., Santana, C. S., Carneiro, P. B. M., Meirelles, A. J., Santos, B. A., Oliveira, A. H. B., Horta, P., Cavalcante, R. M. 2020. “Oil Spill in South Atlantic (Brazil): Environmental and Governmental Disaster.” *Marine Policy* 115 (May): 103879. <https://doi.org/10.1016/j.marpol.2020.103879>.
- Surkes, Sue. 2021. “Marine Monitor Urges Israel to Step up Surveillance or Risk More Oil Spill Woes.” *The Times of Israel*. March 17, 2021. <https://www.timesofisrael.com/marine-monitor-urges-israel-to-step-up-surveillance-or-risk-more-oil-spill-woes/>.
- VisioTerra. 2021. “Baniyas Thermal Power Plant Causes a 1000 Km² Oil Spill, Syria.” <https://www.sentinelvision.eu/gallery/pdf/e16326bf3cd44b2cbf689cd1ad134af2>.
- Voytenko, Mikhail. 2021. “Greek Aframax cleared of spill accusation in Novorossiysk, Black sea.” *FleetMon.com*. August 11, 2021. <https://www.fleetmon.com/maritime-news/2021/34869/greek-afamax-cleared-spill-accusation-novorossiysk/>.
- Wikipedia contributors. 2022. “List of Oil Fields.” *Wikipedia, The Free Encyclopedia*. August 14, 2022. https://en.wikipedia.org/w/index.php?title=List_of_oil_fields&oldid=1104337769.

Georeferenced Multipurpose Real Estate Re-registration Support System

**Bruno G. Batista¹, Vanessa C. O. Souza¹, Melise M. V. Paula¹,
Rafael M. D. Frinhani¹, Bianca R. Bartolomei¹, Lucas Lamounier¹,
Roberto A. Almeida¹**

¹Institute of Mathematics and Computing (IMC)
Institute of Natural Resources (IRN)
Federal University of Itajubá (UNIFEI)
Postal Code 37.500-309 – Itajubá – MG – Brazil

{brunoguazzelli, vanessasouza, melise, frinhani}@unifei.edu.br

{biancarbartolomei, lucaslamounier08, roberto4772}@unifei.edu.br

Abstract. *The implantation and maintenance of a cartographic base is not a simple task, mainly in great territorial extensions, such as for example, Brazil. Cities follow a dynamic system of growth, undergoing significant changes in a relatively short period, which makes urban mapping a major challenge. Outdated cadastral bases make it difficult for managers to take decisions and tax taxes that make up a large part of the municipalities' budget. The objective of this paper is to present a support system for the georeferenced multipurpose real estate re-registration of cities. The system takes into account images collected by drones during overflights in cities. These images are vectorized and the data obtained are compared with the cadastral data provided by the city hall, generating an updated geographic database. In addition, the system is composed of a web application with two modules, Administrative and Citizen, and an application for field collection, Collect Module. Through these modules, managers can analyze the updated geographic database to take strategic decisions about the city and tax taxation fairly. These definitions are in line with Brazilian legislation, following the premises defined in Ordinance number 511 of the Ministry of Cities and in Complementary Law number 101.*

Resumo. *A implantação e manutenção de uma base cartográfica não é uma tarefa simples, principalmente em grandes extensões territoriais, como por exemplo, o Brasil. As cidades obedecem a um sistema dinâmico de crescimento, sofrendo alterações significativas em um período relativamente curto, o que torna o mapeamento urbano um grande desafio. Bases cadastrais desatualizadas dificultam a tomada de decisão dos gestores e a tributação de impostos que compõem grande parte do orçamento dos municípios. O objetivo deste artigo consiste em apresentar um sistema de auxílio para o cadastramento e o recadastramento imobiliário multifinalitário georreferenciado dos municípios. O sistema leva em consideração imagens coletadas por drones durante o sobrevôo nas cidades. Essas imagens são vetorizadas e os dados obtidos são confrontados com os dados cadastrais fornecidos pela prefeitura, gerando uma base de dados geográfica atualizada. Além disso, o sistema é composto por uma*

aplicação web com dois módulos, Administrativo e Cidadão, e uma aplicação para coleta em campo, Módulo de Coleta. Por meio destes módulos, os gestores podem analisar a base de dados geográfica atualizada para a tomada de decisões estratégicas sobre o município e a tributação de impostos de forma justa. Essas definições vão de encontro às premissas definidas na Portaria número 511 do Ministério das Cidades e na Lei Complementar número 101.

1. Introduction

The urban space is dynamic, it changes constantly and, therefore, it is essential to have updated information on the occupation of the territory. This allows governments and the population to have knowledge about the current state and seek to act more effectively and assertively considering the investments in municipalities. However, cities have difficulty maintaining and updating records and inspecting properties, especially the small one, causing a discrepancy in the Municipal Real Estate Registry (CIM – Cadastro Imobiliário Municipal according to Brazilian legislation).

In a simplified way, the Municipal Real Estate Registry is composed of the city's basic plan and the information about different elements of the properties and their owners. An outdated CIM difficulties managers make-decisions actions. This is because managers cannot know, precisely, the portion of their territory and, mainly, its characteristics, such as occupation, territorial distribution, among other elements that must be considered in the formulation of public policies.

In addition, an obsolete CIM impacts the municipality's finances, as it serves as a basis for the calculation of various taxes. In Brazil, Complementary Law number 101 [Brasil 2000] indirectly establishes that all cities must carry out a mapping of their urban area and a detailed evaluation of properties for the purpose of collecting taxes. The cities that do not follow the law can have decreases in revenue collection due to omissions in the updating of the real estate registry. Therefore, the updated registry of urban properties is the main instrument for the correct evaluation of urban service fees and taxes, including the Urban Land and Land Property Tax (IPTU – Imposto sobre a Propriedade Predial e Territorial Urbana according to Brazilian legislation). This update of the CIM occurs through the Real Estate Re-registration.

The objective of this paper consists in developing a support system for the re-registration of properties located in the urban area of the municipalities through geotechnologies. One of the immediate benefits of having an updated Multipurpose Technical Registration and Re-registration (CTM – Cadastramento e Recadastramento Técnico Multifinalitário according to Brazilian legislation) supported by geotechnologies is the identification of the actual real estate situation and the tax capacity of the city. This process is carried out through access to information that composes the municipality's database and that subsidizes the decision-making of managers, as well as the collection of taxes, such as the IPTU, in a fair and accurate way.

For the development of a CTM that meets all these purposes, a system was developed, which involves: (i) imaging the city through UAVs (Unmanned Aerial Vehicles), resulting in an orthophoto¹; (ii) the vectorization of images; (iii) the implementation of

¹Orthophoto is a distortion-free georeferenced aerial photograph. It has geometric accuracy comparable

a geographic database; (iv) collecting, confirming field data and resolving inconsistencies; (v) the data availability in a Geographic Information System (GIS); and (vi) the development of a web application that allows access to information by the population in general, the administration of data by the city hall and access to reports and analytical panels (dashboards) with relevant information for the city decision-makers.

In view of the above, the main contributions of this study are:

- Mapping of the urban area of the municipality and availability of data in CAD, SHAPE and QGIS files;
- The updated geographic database of the city;
- A data API in PostGIS² for integration with other systems;
- Municipal maps such as Neighborhood Map, Sector Map, Plot Map and Generic Value Plan;
- Real estate reporting and notification system for taxpayers;
- Administrative Module, Citizen Module and Collect Module for making maps and information about the property available to society and the city hall.

The remainder of this paper is organized as follows: Section 2 discusses georeferenced multipurpose real estate re-registration; Section 3 presents the methodology applied in this research; the development stages of the proposed system are described in Section 4; finally, in Section 5 the conclusions and premises for future work are discussed.

2. Georeferenced Multipurpose Real Estate Re-registration

According to Ordinance number 511 of the Ministry of Cities [Brasil 2009], the real estate re-registration aims to update the municipal database. In particular, information on taxpayers, public spaces and properties are used. In this context, managers have chosen to carry out the Multipurpose Technical Registration and Re-registration (CTM).

More complete than the CIM, the CTM can be understood as a registration system for the spatial elements that represent the urban structure, consisting of a cadastral base and a georeferenced cartographic base. Georeferencing is understood as the assignment of geographic coordinates to each record in the database, through operations performed in a Geographic Information System (GIS) [Tomlinson 2007]. Thus, the CTM gathers various information that can support municipal management and can be considered a public inventory of all plots of land and properties in a city.

For the registry to become truly multipurpose, it must serve as many users as possible and multiple applications. The CTM development and implementation can benefit from the use of geotechnologies, which have been widely used in the most varied contexts, from the monitoring of environments considered critical, such as solid tailings dams, to the analysis of deforested areas and subject to flooding [da Silva and Fernandes 2020].

Therefore, the georeferenced multipurpose real estate re-registration uses several technologies such as Remote Sensing, the Geographic Information System (GIS) and the Global Positioning System (GPS), and is directly linked to the concept and implementation of Smart Cities. In Smart Cities, a wide variety of Information and Communication

to a topographic map [Habib et al. 2007].

²<https://postgis.net/>

Technology (ICT) solutions are used to deal with various challenges involving socio-economic and administrative issues of a city, seeking to assist the administrative sector in decision-making [Khan et al. 2015] [Xiao et al. 2017].

The criteria used for a city to be considered smart involves projects and actions related to sustainability, nature environment preservation, public spending, mobility, urbanism, technology, communication and innovation [Song et al. 2017] [Allam et al. 2022]. However, decision-making related to these projects and actions depends on defining a modern master plan, as well as updating the municipality's real estate database.

The traditional methodology of real estate re-registration considers the measurement of properties in loco to update their area, which can take months to produce updated maps. Therefore, photogrammetric techniques can be used as an alternative, since they allow high localization precision and fast production of data [Fraser 2001] [Maas and Hampel 2006] [KARATAŞ and ALTINIŞIK 2020]. Regardless of the technique used, the legal limits of the properties must be obtained with adequate precision to guarantee the correctness of the property location and of the area [Brasil 2009].

In re-registration projects supported by high spatial and geometric resolution images, such as orthophotos, the main process is called vectorization. In vectorization, analysts use specialized software to manually contour the roofs identified in the images in order to generate geometries, called polygons, which represent the buildings [Habib et al. 2007]. In this case, the property area corresponds to the projection area of the roof. However, this process can be a bottleneck in re-registration projects because it is manual and depends on the analyst's experience.

Automatic image classification is an alternative solution to manual vectorization. Therefore, several authors have applied this technique to identify roofs in urban areas [Chen et al. 2018] [Pilinja Subrahmanya et al. 2021] [Trevisiol et al. 2022] [Jia et al. 2022]. Automatic classification tends to minimize the effort of manual vectoring and shorten the time to obtain results. However, depending on the method used, the accuracy can be compromised, impacting the representation of the analyzed area.

3. Methodology

For real estate re-registration purposes, it is necessary to use Very High Resolution (VHR) spatial images. Spatial resolution refers to the area that an image pixel represents on the ground. In this case, each pixel is equivalent to an area between 0.3 and 1 meter on the ground [Shao et al. 2021]. According to [Myint and Stow 2011], image classifications using object-based approaches are more applicable to high spatial resolution data, where the objects of interest are larger than the ground resolution element.

There are several works in the literature that propose different ways of recording urban or rural areas through geotechnologies, whether through satellite images, topography or aerophotogrammetry [Moraes et al. 2013] [Crommelinck et al. 2017] [Fetai et al. 2019] [Cienciała et al. 2021] [Fetai et al. 2021] [Ajayi and Oruma 2022].

Although there are studies with a main focus on real estate re-registration, no studies were found that propose a system to assist the georeferenced multipurpose real estate re-registration in cities, in order to help in the decision-making process by managers, in the taxes calculation and in the visualization of information by stakeholders and citizens.

For this, the updating of the buildings area in this paper is based on the concept stipulated by the standard NBR 1272:2006 of ABNT (Brazilian Association of Technical Standards)³, which defines that the built-up area corresponds to “the total covered area of a building, which includes the projection area of the building’s roof”.

Following this concept, the study carried out in this paper assumes that the buildings roof area is the essential information to obtain a better approximation of the real built area. However, considering that the roof can have eaves and that the building can have more than one floor, the built area generated by this study is calculated according to Equation 1:

$$\text{builtArea} = (\text{vectorizedArea} - \text{eavesArea}) \times \text{floorsNumber} \quad (1)$$

Figure 1 presents an overview of the proposed system, described in more details in Section 4.

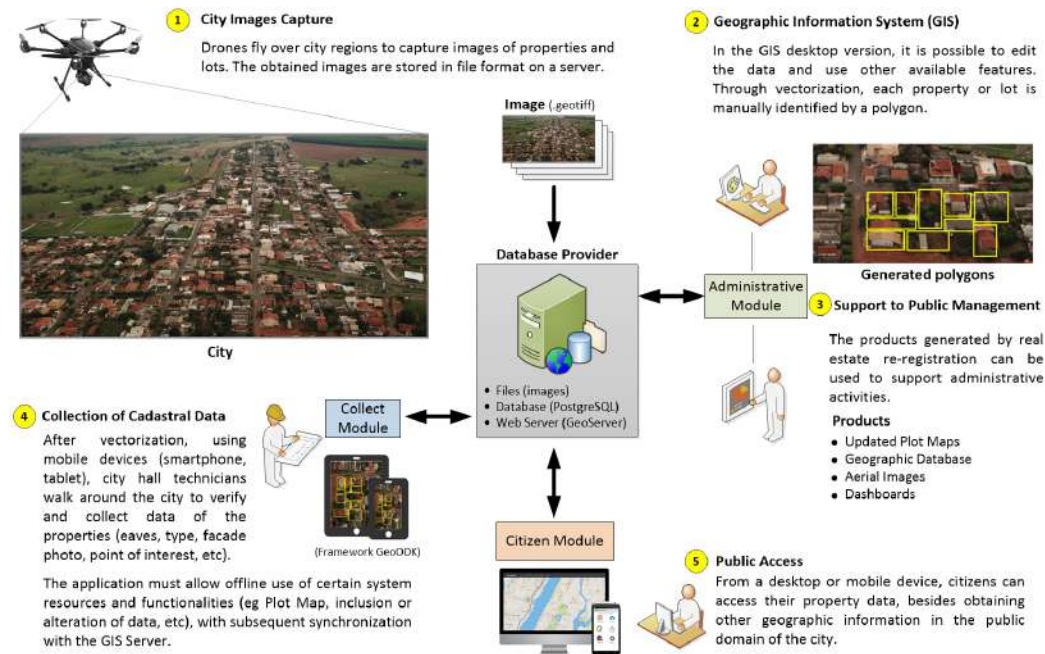


Figura 1. System proposed to support the georeferenced multipurpose real estate re-registration.

4. Proposed System

For the development of the georeferenced multipurpose real estate re-registration support system proposed in this paper, it is necessary: (i) to capture images of the city through UAVs, generating the city orthophoto; (ii) the images vectorization; (iii) the implementation of a geographic database; (iv) to collect, confirm data in the field and resolve inconsistencies; (v) to make these data available in a geographic information system; and (vi) the development of a web application that allows access to information by the population

³ Available at: <http://www.cub.org.br/abnt-nbr-12721-2006>

in general, the administration of data by the city hall and access to reports and analytical panels (dashboards) with relevant information for the city decision-makers. Figure 2 presents the architecture of the developed system.

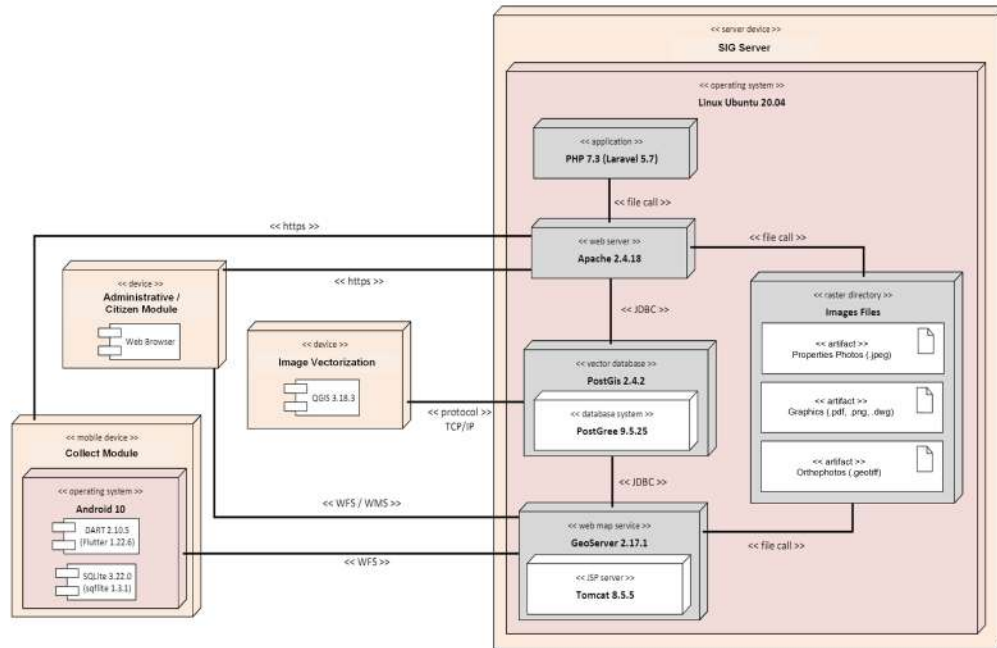


Figure 2. Proposed system architecture.

4.1. Phase 1 - Images capturing

The city images are captured by drones, in a period previously defined and disclosed to the population and other agents, such as firefighters and police officers. For imaging it is important that the drone is equipped with a high resolution camera, so that there is a high overlap of the images. This decreases the distortions on roofs in the orthophoto and increases the accuracy during the images vectorization.

The captured images will support the execution of the real estate re-registration of the city from the identification of public places, blocks, lots, real estate units and improvements. Therefore, Phase 1 consists of:

- Carrying out a photogrammetric aerial survey of the city;
- Processing the data obtained in the aerial survey, including the collection and processing of control points using the Real-Time Kinematic Global Navigation Satellite System (RTK GNSS) [Roosevelt 2014];
- Generating a point cloud. A point cloud is a set of georeferenced points capable of recreating areas and structures. These points are obtained through drone mapping and image processing, and allow obtaining various information about the area such as depth, elevation, geometry and location [Rosnell and Honkavaara 2012];
- The generation of the digital model of the terrain, known as Topographic Plan;

- The generation of rectified orthomosaic and;
- The compilation of the rectified orthomosaic with a spatial resolution of up to 5 cm per pixel in plan (in the study carried out in this paper). However, in the study presented in [Arantes et al. 2018], the measurement values, obtained through aerial images, show that the areas built without eaves and with eaves do not vary significantly between the spatial resolution of 2, 5 and 10 centimeters per pixel.

4.2. Phase 2 - Images vectorization

In the image vectorization phase, an orthophoto captured by a drone serves as a basis for the delimitation of spatial objects in the city, with a spatial resolution of less than or equal to 5 centimeters per pixel. As mentioned, the built-up area is calculated from Equation 1. Streets, axes, blocks, lots, real estate units, urban hydrography and other improvements are vectorized. The delimitation of lots and blocks are made from the visualization of their physical limits in the orthophoto, such as walls, fences and sidewalks. In the absence of physical limits, the delimitation can be made from the measures established in the plot map, provided by the city hall.

As a result of vectorization, the vectorized object is outlined based on the orthophoto, in a way the roof structured geometry is converted into a vector representation, a polygon. The polygon goes through a labeling process, in which a label is assigned to identify it in relation to the city, with information like the district, sector, block, land lot and real estate unit.

During images vectorization phase, it is common to find adversities arising from comparative analyzes between the data provided by the city hall, through the plot map, and the data captured by the drone. There are three main types of adversities:

- Remembrance and Dismemberment: land lots that were merged or divided, but did not have their records updated in the system;
- Area invasion: lots invading institutional or not subdivided areas;
- Lots and real estate units that are not included in the city hall's database: they do not have records at the city hall, but they were identified during the images capturing.

When an adversity is identified, it must be presented to the city hall for resolution and subsequent adjustment in the image vectorization. In the absence of adversities, vectorized data and cadastral data are stored in the Geographic Database. Figure 3 presents the described process.

4.3. Phase 3 - Geographic database

In the geographic database phase, the cadastral data about the properties and public places in the city are associated with vectorized geometries. A data model is generated, which stores both descriptive attributes and geometries, using the PostGIS Database Management System.

For this step to be carried out, the city hall must provide the cadastral data of all properties, called Cadastral Information Bulletin (BIC - Boletim de Informações Cadastrais according to Brazilian legislation) and the cartographic plant of the city (plot map), whether georeferenced and digital or not. In case it is not georeferenced, the city hall

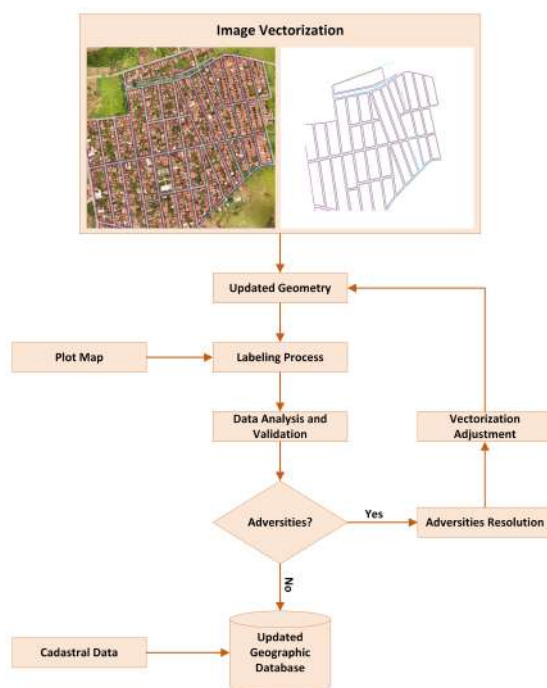


Figura 3. Georeferenced multipurpose real estate re-registration process.

must deliver the plot map with the lots identified by the inscription sector-block-lot and not just with the subdivision number. In addition, the city hall must provide an updated list of all public places.

The lack of information makes it difficult to map between the records present in the municipal database and the images captured in Phase 1 and vectorized in Phase 2. Therefore, three validation processes are carried out: labeling, geometric and cadastral. In the labeling validation, it is verified if the labels assigned to the vectorized images are correct and represent the real location of the vectorized object in the city. In the geometric validation, it is verified if the generated polygons follow the geometry present in the plan provided by the city hall. Finally, in the cadastral validation, it is verified whether the data generated from Phases 1 and 2 match the data provided by the city hall. There may be adversities such as different names for the same street, for example.

After going through the labeling, geometric and cadastral validations, in case of positive validations, the record is inserted into the database, or is sent to an adversity database for validations with the city hall. Figure 3 presents the process for analyzing and validating the data that will compose the updated geographic database and be accessed by developed applications.

4.4. Phase 4 - Field data collection

In the field data collection phase, people selected by the city hall and duly identified, known as collectors, walk through the streets of the city with devices such as tablets and/or smartphones, collecting and verifying information about the properties using the Collect Module.

Collectors do not enter the property. They access the data available in the geographic database via some mobile device running Collect Module, verify if these data correspond to what they see during the in-loco visit and, if necessary, update the property's registry information. Information about eaves and number of floors of the property are also collected for all covered areas that compose the real estate unit, as well as photos are taken at the time of the visit, in front of the property. There is no interaction with the owners. In this way, the data contained in the property Cadastral Information Bulletin (BIC) are updated through a mobile application developed for this purpose.

4.5. Phase 5 - Results Availability

In the results availability phase, a web application with two modules was developed. The first, called Administrative Module, allows the application to support the re-registration project, so that the city hall team can update property registration data and resolve adversities detected between what is visualized in the orthophoto and in the cartographic plan of the city. In addition, the Administrative Module provides operational and strategic reports on the database and an API (Application Programming Interface) for data sharing, which are crucial for decision-making by the city administrative sector.

The second module, called Citizen Module, allows citizens to consult property data through a SIG-WEB (Geographic Information System) interface. This interface displays the city map with vectorized data layers (axis, lot, block, covered area, among others). The citizen can browse the map and obtain basic information about the properties in three different ways: by the property registration number (BIC), by the reduced number of the BIC, and by the address. It is important to highlight that only data considered public by the city hall are available for consultation, which cannot be edited through the Citizen Module. Editing these data can only be performed by employees authorized by the city hall through the Administrative Module, which requires login.

Finally, in addition to the web application, the final data are structured in a Geographic Information System, connected directly to the geographic database. At the end, the cartographic plant of the city is updated and made available in GIS⁴ and CAD formats⁵, as well as the Geographic Database, with updated registration data. All the services are implemented in the cloud.

5. Conclusion

Cadastral data about the territory occupation are essential for knowledge, management and strategic decision-making by the administrative sector. In large territorial areas, such as Brazil, the implementation and maintenance of a geographic database with city property records is not a trivial task.

When dealing with urban mapping, the issue is delicate, because cities follow a dynamic system of growth, undergoing significant changes in a relatively short period. For this reason, it is extremely important that municipal managers follow the Ordinance number 511, of the Ministry of Cities, and Complementary Law number 101. Their premises are important in order to carry out the mapping of the city's urban area and

⁴<https://qgis.org/en/site/>

⁵<https://www.autodesk.com.br/solutions/cad-software>

properties detailed evaluation for the purpose of collecting taxes and strategic decision-making related to public management.

The objective of this paper is to present a support system for the georeferenced multipurpose real estate re-registration of cities. The study developed fits as a process technological innovation project, since it incorporates new technologies to the real estate re-registration process using images of unmanned aerial vehicles. The insertion of two software products stands out: a Web Application, composed by Administrative and Citizen Modules, and a Collect Module that support the solution of adversities found in the vectorization stage and the collection of data in the field, essential for real estate re-registration. This paper also innovates by applying changes in the methodology for executing the real estate re-registration process, since it defines rules and techniques for vectorization, for updating the geographic database and calculating the area of the properties.

Usability and performance analyses will be carried out in future works, taking into account the loading of data and orthophotos in the Administrative, Citizen and Collect Modules.

Referências

- Ajayi, O. G. and Oruma, E. (2022). On the applicability of integrated uav photogrammetry and automatic feature extraction for cadastral mapping. *Advances in Geodesy and Geoinformation*, pages e19–e19.
- Allam, Z., Sharifi, A., Bibri, S. E., and Chabaud, D. (2022). Emerging trends and knowledge structures of smart urban governance. *Sustainability*, 14(9):5275.
- Arantes, B. H. T., Arantes, L. T., Ventura, M. V. A., Costa, E. M., Bessa, M. M., Baliza, L. M., and Moraes, V. H. (2018). Uso de drones na atualização de área construída de imóveis urbanos. *Scientia Plena*, 14(10).
- Brasil (2000). Lei nº 101, de 4 de maio de 2000. estabelece normas de finanças públicas voltadas para a responsabilidade na gestão fiscal e dá outras providências. diário oficial da república federativa do brasil.
- Brasil (2009). Portaria nº 511, de 7 de dezembro de 2009. estabelece diretrizes para a criação, instituição e atualização do cadastro territorial multifinalitário (ctm) nos municípios brasileiros. diário oficial da república federativa do brasil.
- Chen, R., Li, X., and Li, J. (2018). Object-based features for house detection from rgb high-resolution images. *Remote Sensing*, 10(3):451.
- Cienciała, A., Sobolewska-Mikulska, K., and Sobura, S. (2021). Credibility of the cadastral data on land use and the methodology for their verification and update. *Land Use Policy*, 102:105204.
- Crommelinck, S., Bennett, R., Gerke, M., Yang, M. Y., and Vosselman, G. (2017). Contour detection for uav-based cadastral mapping. *Remote sensing*, 9(2):171.
- da Silva, A. O. and Fernandes, R. A. S. (2020). Smart governance based on multipurpose territorial cadastre and geographic information system: An analysis of geoinformation, transparency and collaborative participation for brazilian capitals. *Land Use Policy*, 97:104752.

- Fetai, B., Oštir, K., Kosmatin Fras, M., and Lisec, A. (2019). Extraction of visible boundaries for cadastral mapping based on uav imagery. *Remote sensing*, 11(13):1510.
- Fetai, B., Račić, M., and Lisec, A. (2021). Deep learning for detection of visible land boundaries from uav imagery. *Remote Sensing*, 13(11):2077.
- Fraser, C. S. (2001). Photogrammetric camera component calibration: A review of analytical techniques. *Calibration and Orientation of cameras in Computer Vision*, pages 95–121.
- Habib, A. F., Kim, E.-M., and Kim, C.-J. (2007). New methodologies for true orthophoto generation. *Photogrammetric Engineering & Remote Sensing*, 73(1):25–36.
- Jia, J., Cui, W., and Liu, J. (2022). Urban catchment-scale blue-green-gray infrastructure classification with unmanned aerial vehicle images and machine learning algorithms. *Frontiers in Environmental Science*, page 734.
- KARATAŞ, K. and ALTINIŞIK, N. S. (2020). The effect of uav usage on detail points in cadastre update studies: Çorum-karaköy case study. *International Journal of Environment and Geoinformatics*, 7(2):140–146.
- Khan, Z., Anjum, A., Soomro, K., and Tahir, M. A. (2015). Towards cloud based big data analytics for smart future cities. *Journal of Cloud Computing*, 4(1):1–11.
- Maas, H.-G. and Hampel, U. (2006). Photogrammetric techniques in civil engineering material testing and structure monitoring. *Photogrammetric Engineering & Remote Sensing*, 72(1):39–45.
- Moraes, I. C., Correa, E. A., Da Conceicao, F. T., and Lucimari Ap F, G. R. (2013). Geotechnology applied to land use and sealing in urban areas. In *Joint Urban Remote Sensing Event 2013*, pages 099–102. IEEE.
- Myint, S. and Stow, D. (2011). An object-oriented pattern recognition approach for urban classification. In *Urban Remote Sensing: Monitoring, Synthesis and Modeling in the Urban Environment*, pages 129–140. Wiley-Blackwell.
- Pilinja Subrahmanya, P., Haridas Aithal, B., and Mitra, S. (2021). Automatic extraction of buildings from uav-based imagery using artificial neural networks. *Journal of the Indian Society of Remote Sensing*, 49(3):681–687.
- Roosevelt, C. H. (2014). Mapping site-level microtopography with real-time kinematic global navigation satellite systems (rtk gnss) and unmanned aerial vehicle photogrammetry (uavp). *Open Archaeology*, 1(1).
- Rosnell, T. and Honkavaara, E. (2012). Point cloud generation from aerial image data acquired by a quadrocopter type micro unmanned aerial vehicle and a digital still camera. *Sensors*, 12(1):453–480.
- Shao, H., Song, P., Mu, B., Tian, G., Chen, Q., He, R., and Kim, G. (2021). Assessing city-scale green roof development potential using unmanned aerial vehicle (uav) imagery. *Urban Forestry & Urban Greening*, 57:126954.
- Song, H., Srinivasan, R., Sookoor, T., and Jeschke, S. (2017). *Smart cities: foundations, principles, and applications*. John Wiley & Sons.

- Tomlinson, R. F. (2007). *Thinking about GIS: geographic information system planning for managers*, volume 1. ESRI, Inc.
- Trevisiol, F., Lambertini, A., Franci, F., and Mandanici, E. (2022). An object-oriented approach to the classification of roofing materials using very high-resolution satellite stereo-pairs. *Remote Sensing*, 14(4):849.
- Xiao, C., Chen, N., Gong, J., Wang, W., Hu, C., and Chen, Z. (2017). Event-driven distributed information resource-focusing service for emergency response in smart city with cyber-physical infrastructures. *ISPRS International Journal of Geo-Information*, 6(8):251.

A QGIS plugin for BONDS project: integrating field data with geographical, remote sensing and health information

Rogério Flores Júnior^{1,2}, Vitor Conrado Faria Gomes^{1,3},
Karine Reis Ferreira¹, Gilberto Ribeiro de Queiroz¹, Marie-Paule Bonnet²

¹Image Processing Division, National Institute for Space Research (INPE)
São José dos Campos – SP – Brazil

²Environment and Resources Department – Institut de recherche pour le développement (IRD)
Montpellier – Occitanie – France

³C4ISR Division – Institute for Advanced Studies (IEAv)
São José dos Campos – SP – Brazil

rogerio.junior@inpe.br, vitor@ieav.cta.br

gilberto.queiroz@inpe.br, karine.ferreira@inpe.br

Abstract. *Open source Geographic Information System (GIS) have been fostering spatial data research such as Earth observation and environmental monitoring for more than 30 years. More recently, globally available geospatial information combined with web technologies are providing new environments and tools for data handling. Thus, binding the mapping and processing capabilities of traditional GIS to the accessibility and reliability of web-based data providers can bring new opportunities for research. In this paper, we built a QGIS plugin to explore the integration of different public data providers in Brazil along with field data produced by the BONDS project. The biODiversity conservation with Development in Amazon wetlands project (BONDS) proposes to develop biodiversity scenarios for the Amazonian floodplains aiming to support solutions to preserve biodiversity and ecosystem services. The use of web services enabled dynamic and fast access to several products ranging from remote sensing images, land use and land cover, territorial cartography, water quality, to COVID-19 health data, and more.*

1. Introduction

Within the technological revolution of the last decades, the advancement of the internet and mobile communication changed how information is produced, stored, and consumed [Tole et al. 2013]. With these new technologies and the possibility of a fast and reliable Global Navigation Satellite System (GNSS), data is now produced on the fly and mostly with Geographical Information (GI) attached to it [Maguire and Longley 2005]. Although this massive amount of data can be portrayed as big data, its geographical features bring new paradigms [Lee and Kang 2015]. Nowadays, GI data has become a paramount framework to sample and represent phenomena and patterns of the real world in its digital twin. In this matter, Geographical Information Systems (GIS) comes into the picture to meet the necessity to understand, represent and forecast data by providing a geographic ecosystem with tools and techniques to visualize, analyze and model the GI data [Raju 2006].

In the internet era, all areas of knowledge including GIS have been influenced to a large extent by internet standards and technologies, resulting in the increasing availability of online applications using distributed web services. Ready-to-use services are shaping how we interact with data by providing standards and tools to discover, access and acquire data in an organized, fast and reliable way. [Lemmens 2006]. There is apparent migration of traditional GIS tools to web based solutions, however, these approaches still suffer with paradigms such as the need for uninterrupted and reliable internet, poor mapping and visualization features, lack of advanced processing tools, and lack of computational power for everyone. Therefore, the use of desktop or client-side GIS tools are very likely to continue to be needed [Michaelis and Ames 2012].

In spite of its benefits, traditional desktop based GIS are burdened with data duplicity, outdated data sources, manual data input, and the inability to easily share data. Therefore, combining client-side GIS with web based data tools may enhance the user's experience. The use of well established practices such as the Open Geospatial Consortium (OGC) standards could be useful for promoting better data access, normalization and shareability. These new services can be very helpful to the scientific community as research projects become more globalized. An example is the *Balancing biODiversity conservation with Development in Amazon wetlandS* (BONDS) an international research project comprising seven countries and funded by the joint call of the Belmont Forum and BiodivERsA, within the scope of the BiodivScen ERA-Net COFUND program. This project proposes to develop biodiversity scenarios for the Amazonian floodplains, combining conservation and development issues. Thus, it aims to find, along with stakeholders, solutions to preserve biodiversity and ecosystem services in the context of socioeconomics and climate change.

BONDS research groups are actively generating Earth Observation (EO) derived products based on multi-disciplinary GI data which, despite being open and freely available, are often decentralized and hard to reach. In addition, there is no single place of management and distribution for data produced by the project. With the project's research areas diversity, some collaborators may not have the access or computational ability to search and effectively consume the data in the present conditions. Therefore, this work aims to explore data integration methods and evaluate the benefits and drawbacks of current technologies. This paper presents a QGIS plugin to provide access to data from both BONDS project and Brazil's public and freely available datasets. The plugin was designed to abstract the computational technologies to aid BONDS researchers and external users to reach as much relevant data as possible.

2. Materials and methods

2.1. BONDS Study area

The study areas for the BONDS project are floodplain areas along the Brazilian Amazon rain forest. Those areas encompass the high, medium and low Amazon river regions, and are located in both Amazonas and Pará states. The Curuai lake floodplains are the most important BONDS site and also the focus of this paper (Figure 1). The Lake is located at the lower Amazon river region ($-55.489^{\circ}W$, $-2.111^{\circ}S$), next to the cities of Santarém and Óbidos, being around 900km from the Amazon river delta. Figure 1 presents a few

¹<http://www.bonds-amazonia.org/>

QGIS layers obtained with the BONDS plugin. These data are the CREN Land Use and Land Cover (LULC) for Pará state (2018), CGEO federal water bodies, BONDS water quality field data, and a true color Landsat-8 image from a 16 day (08/28-09/12/2020) datacube from Brazil Data Cube (BDC) project.

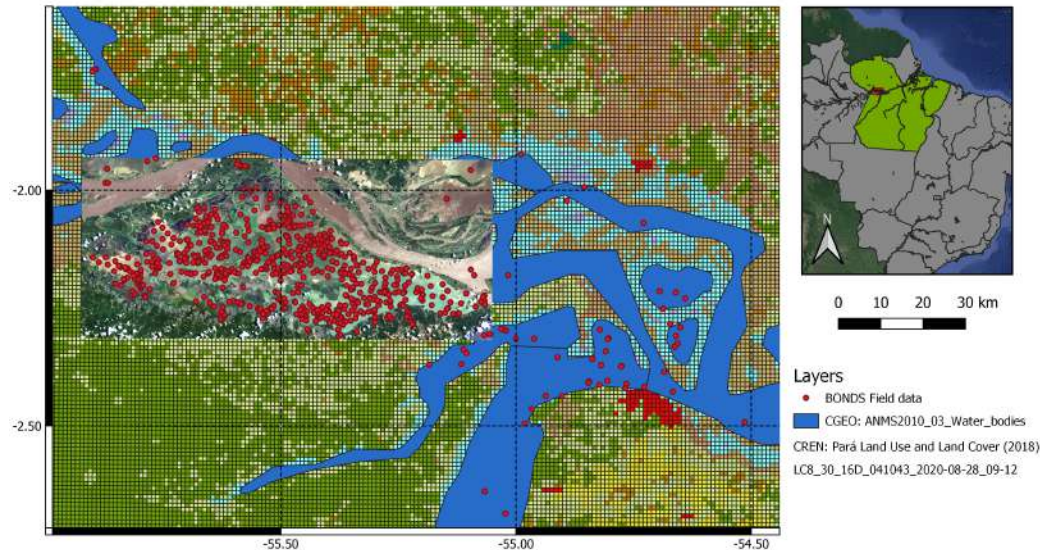


Figure 1. BONDS main study area with layers obtained through BONDS plugin. In the background CREN LULC for Pará state (2018). As vectors the CGEO federal water masses and BONDS water quality field data. On top a true color RGB Landsat-8 image from a 16 day datacube from BDC.

2.2. Brazilian public open data

Although policies regarding the use and access of public data have been discussed since the 1980s, the topic started to grow in Brazil in the early 2010s with initiatives for boosting government's transparency. Brazil's federal open data policy Decree 8.777/2016 established rules and guidelines for public agencies to comply with the Open Government Data plan (OGD) for public data sharing [Silva and Pinheiro 2019]. Surveying public agencies identifying compliance to OGD [Silva and Pinheiro 2019] conclude that although progress is slow, agencies are moving towards open datasets through Application Programming Interface (API) services.

2.2.1. Brazilian National Data Infrastructure (INDE)

The INDE proposes an integrated set of technologies, standards, policies, monitoring mechanisms and procedures to centralize data information of several governmental providers, ranging from universities and public agencies to research facilities. It is composed of 60 institutions at academic, municipal, state or national levels such as National Institute for Space Research (INPE), Amazon Protection System (SIPAM), Brazilian Institute of the Environment and Renewable Natural Resources (IBAMA), National Water

Agency (ANA), and Brazilian Institute of Geography and Statistics (IBGE). INDE's catalog comprises data from territorial cartography, LULC, biodiversity, water related products and many more.

INDE promotes common standards and agreements to facilitate and standardize data generation, storage, access, sharing, and dissemination. Thus, any partner institutions must comply with OGC standards and guidelines to be able to provide geoservices through INDE's infrastructure [INDE 2022]. Geoservices are web services specific to the geospatial domain, constituting a powerful set of functionalities to collect, store, retrieve, transform and display spatial data. These functionalities can be accessed through web browsers or any other application as programming languages and GIS.

The OGC is an international non-profit organization for creating spatial data sharing standards allowing government agencies, NGOs, companies, academic and research entities to make their geospatial data openly available [Baumann 2010, Giuliani et al. 2013]. Today, OGC is composed of more than 500 organizations and has 58 common open source standards [OGC 2022]. All specifications intend to support data interoperability by following a common architecture [Baumann 2010]. The main services at OGC for EO are: a) Web Map Service (WMS); b) Web Feature Service (WFS); c) Web Coverage Service (WCS); d) Web Processing Service (WPS); e) Web Coverage Processing Service (WCPS); and f) Catalog Service for Web (CSW).

The INDE specification is based on the OGC WMS, WFS, and WCS geoservices [CONCAR 2010]. The WMS is a protocol to serve georeferenced maps optimized for visualization, allowing users to view and consume its information but not meant for modeling or processing [Baumann 2010]. The WFS protocol, on the other hand, defines a set of interfaces for accessing geographic information as feature and feature properties. In short, WFS allows users to access a geographic phenomenon represented in vector format [Giuliani et al. 2013]. Also, WFS provides query capabilities that can be based on spatial and non-spatial constraints. Similar to WFS, the WCS provides datasets in matrix or raster format, however, it is used to represent phenomena with continuous spatial-temporal variations called *coverages*.

2.2.2. Remote sensing

In terms of remote sensing images, Brazil is a historical reference of open public data being one of the first nations to freely open its satellite catalog [Santos 2005]. More recently, the BDC project from INPE is producing and distributing data cubes of satellite images. It is a research, development and technological innovation project to produce, store, manage and analyze cubes of medium resolution satellite data [Ferreira et al. 2020]. The BDC image catalog follows the Spatio Temporal Asset Catalogs (STAC) specification providing a standard interface that allows users and applications to query and discover all image collections and data cubes stored in the project databases. To support and abstract the STAC services provided, the BDC team developed packages in python (stac.py) and R (rstack). The project also provides a web portal where users can access, visualize, and download images and data cubes [Ferreira et al. 2020].

²<https://brazildatacube.dpi.inpe.br/>

The project is also currently developing computational environments and tools with cutting-edge technologies for dealing with EO data. The project provides geoservices using common standards such as Tile Map Service (TMS) by the Open Source Geospatial Foundation (OSGeo) and WMS, WFS, and WCS by the OGC and ISO [Ferreira et al. 2020]. Also, at the assets level (image bands) the raster files are distributed in a Cloud Optimized GeoTIFF (COG), allowing for dynamic access to image subsets.

There are several public providers producing and sharing EO derived products, such as the Modular system for continuous monitoring of inland waters (MAPAQUALI) and Monitoring of Deforestation of the Brazilian Amazon Forest by Satellite (PRODES) projects. MAPAQUALI is an experimental platform for monitoring water quality parameters through satellite images [de Lucia Lobo et al. 2017]. PRODES carries out satellite monitoring of clear-cut deforestation in the Legal Amazon and, since 1988, has produced annual deforestation rates in the region, aiming to establish public policies.

2.2.3. Health COVID-19

The COVID-19 global pandemic was catastrophic for mankind, resulting in millions of lives lost and a global economic recession. Its fast and easy spread worldwide brought to light the need for disease monitoring and control in a globalized society. One of the first data portals to collect and monitor COVID-19 was the Johns Hopkins Coronavirus Resource Center (CRC) [Dong et al. 2020]. It was considered the “*go-to data source*” for COVID-19 by TIME magazine in 2020. The CRC aim is to collect data available on cases, deaths, tests, hospitalizations, and vaccines worldwide.

Brazil has a robust universal healthcare system the Unified Health System (SUS) which comprises several services at the federal, state and municipal levels. SUS provides essential, universal, and free access to health-related services for the whole country. More than 80% of Brazil’s population rely on SUS [Stopa et al. 2017]. To handle its massive digital infrastructure the SUS Information Technology Department (DataSUS) was created in 1991. DataSUS developed over 200 systems and is responsible for organizing and providing free access to the data generated by SUS. At the OpendataSUS portal several public health datasets are available [BRASIL 2022b, BRASIL 2022a]. This data is being organized following the *Comprehensive Knowledge Archive Network* (CKAN) standards allowing users to send HTTP requisitions to retrieve the data. However, not every dataset in the portal is already ready for the CKAN API search. Currently, at the portal there are two datasets through API search, the COVID-19 vaccination status and flu syndrome of suspected COVID-19 cases, other datasets are available in Comma Separated Values (CSV) for direct download.

2.2.4. BONDS field data

The field data used in this work is composed by Instrumentation Laboratory for Aquatic Systems (LabISA)³ Amazon datasets for water quality parameters at BONDS scope. This data is in a PostgreSQL database with PostGIS extension as spatial tabular structures

³<http://www.dpi.inpe.br/labisa/>

due to their relational dependence. An access at LabISA's database allowed for direct communication with PostgreSQL.

2.3. Plugin architecture

In order to develop the plugin we applied some technologies and standards of web services. Web services are a way to expose the functionality of an information system and make it available through standard web technologies [Gottschalk et al. 2002]. These definitions may vary as these services are not single straightforward concepts but a complex pile of technological layers with common communication languages. To develop the plugin, three layers are necessary Data, Service and Client.

The Data layer is related to the data storage, organization and management. This layer is the farthest from the users and is totally abstracted from them, especially in the case of cloud storage solutions. It is tied to each provider (data owner or manager) and their choice of standards and technologies. At this layer, we can have data management with a traditional Database Management System (DBMS) (e.g. PostgreSQL [Momjian 2001]). On other hand, data can also be stored as plain files and distributed with a server manager, such as the Geoserver.

The Service layer is composed of all the software necessary to provide a functional environment to access the data and its information. This layer allows users to consume data without knowledge of internal data storage infrastructures through a set of APIs. These APIs can be standardized with communication standard protocols to ensure interoperability between applications. This allows the services to be language and technology independent.

The Client layer is built to consume the services and provide means of abstraction from the code and is the closest to the end-user. Clients are developed to use the API provided at the server layer, allowing for several kinds of applications to be built ranging from back-end to front-end, being independent of the programming language. For example, languages such as Python, R, PHP, and C++ can consume services for back-end operations and JavaScript can be used for front-end applications. For geospatial data, these clients often provide a Graphical User Interface (GUI) for mapping and several tools for data analyses, for example, the QGIS and ArcGIS. More recently, with the maturing of the internet, web-based applications consuming web services and providing interfaces and tools are becoming widespread.

Figure 2 presents the BONDS QGIS plugin service architecture in layers related to each of the layers presented before. At the bottom are the data providers by area of interest, in the middle are the services used for each, and on top is the BONDS QGIS plugin as a client.

2.4. QGIS Plugin development

The development of the plugin interface and code was carried out using a conjunction of three technologies: Python, QGIS, and QT Designer. This project uses, Python version 3.9.0 as available within QGIS components. QGIS is a free and open-source cross-platform desktop GIS application that supports viewing, editing, printing, and analysis of geospatial data [Lawhead 2017]. For QGIS, the selected version was the 3.24.3 Tisler.

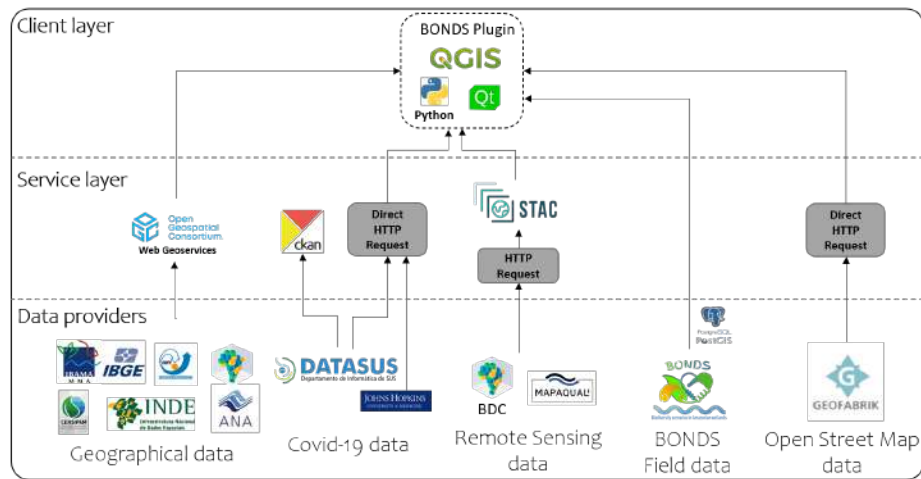


Figure 2. BONDS plugin architecture.

QGIS comes with a distribution of QT Designer by default. QT Designer is an integrated development environment that simplifies GUI application development.

3. Results and discussion

To understand the data available in the context of BONDS, it was conducted a survey of possible data providers and products on five topics of interest: remote sensing imagery and products, field remote sensing, laboratory data, territorial cartography, and health data related to COVID-19. Table 1 presents the number of providers, relevant products and achievable products found. The relevant products were selected by assessing BONDS' objectives to develop studies related to biodiversity, aquatic ecosystems, climate change, and health. The achievable products are the ones that during the plugin implementation were possible to be accessed and consumed. Data providers and products will be presented in the following sections.

Topics of interest	Number of providers	Relevant products	Achievable products
Remote Sensing Imagery	2	10	4
Laboratory data	1	7	7
Field remote sensing	1	3	3
Territorial Cartography	7	25	25
Health COVID-19	3	7+	6

Table 1. Data availability by area of interest, BONDS relevant products and achievable products.

The next step was to catalog the maximum information of each provider and product to identify each group of technology and build the plugin tools. Therefore, the results obtained in this paper will be presented following the BONDS plugin development, as each tab represents a different data as well as a technology. This aims to simplify the access of different kinds of data and to make possible to evaluate the technologies applied to store, catalog and distribute it. The available tabs are: Intro, Geoservices, Remote sensing, BONDS field data, Health COVID-19, and Utils.

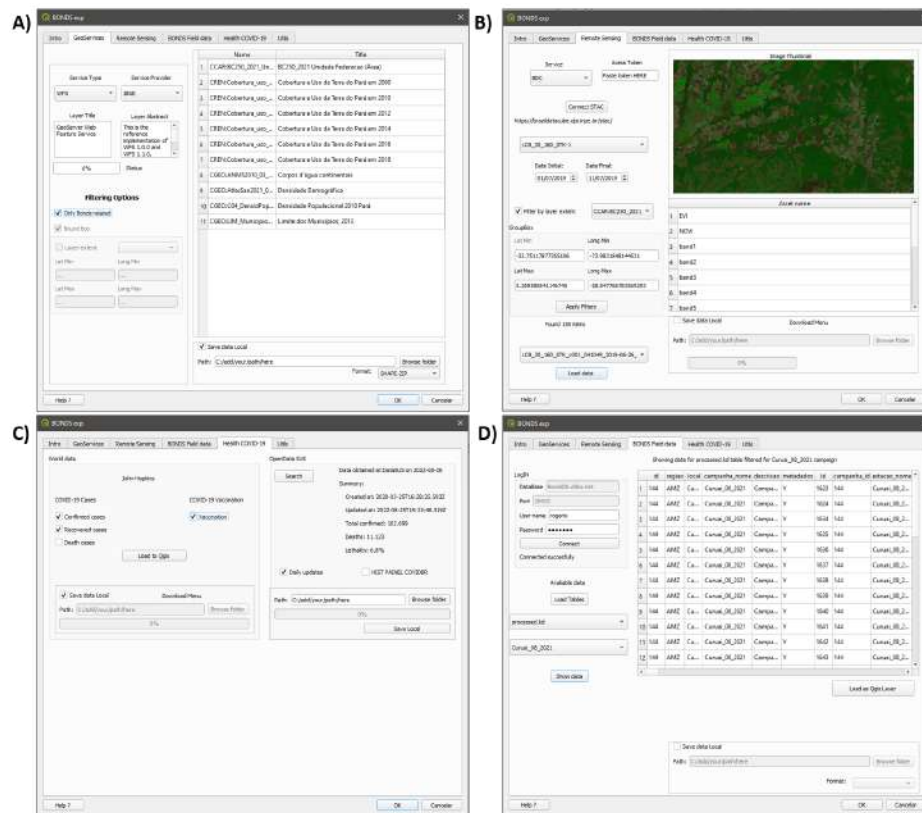


Figure 3. BONDS QGIS plugin main tabs: a) Geoservices; b) Remote sensing; c) BONDS field data; d) Health COVID-19.

3.1. Geoservices

The main Geoservice explored in this paper was the WFS. In total, there are 11 possible operations to WFS requisitions. At the plugin we have focused mainly on the discovery and query operations. Discovery allows the service to be interrogated to determine its capabilities and to retrieve the application schema that defines the feature types that the service offers. Query allows features or values of feature properties to be retrieved from the underlying data store based upon constraints, defined by the client, on feature properties [Michaelis and Ames 2012].

Through the data survey, it was possible to identify seven providers with Geoservice enabled datasets. These providers are: INDE, BDC, INPE, SIPAM, IBAMA, ANA, IBGE. We selected a list of 25 products (layers) that could be of BONDS direct interest. However, at the GUI the user is capable of removing this filter and browsing all the provider's data. It also has the option to filter by bounding box using the WFS capabilities to send parameters along the requisitions (Figure 3a). After the filters choice, WFS vector layers are displayed and made available to be loaded to QGIS or downloaded.

With the WFS capabilities we were able to consume data from several providers in a single standardized format. This brings source and data reliability to the plugin as more and more providers join the standards. Also, allied with newer client technologies, the access and consumption of data were very fast when compared with traditional

downloads. However, on the negative side, we experienced some hiccups with the layer's dynamic load. First, with instability and latency both from local or server, the WFS requisitions often retrieved "timeout error" as the time limit for a response was reached. Also, with large datasets such as the LULC data, the loading was slower and more susceptible to interruptions, thus breaking the data already loaded. The use of resilient layers with microservices can offer solutions to reduce these problems [Nascimento et al. 2020].

From the data point of view, challenges for data integration such as temporal differences, geodetic datum, data format, and semantic inconsistencies were expected [Li et al. 2004, Li 2008]. In our dataset, temporal differences could be seen in data acquired or produced with different time spans, requiring the users to be aware of the date information. Datum problems were observed with data in different coordinate systems and projections, bringing spatial compatibility issues. Semantic problems were mainly related to different nomenclature or class systems among products. Data format was less of a problem as GDAL and QGIS offers many tools to ingest the data, however, data heterogeneity among providers is still an issue.

3.2. Remote sensing

For the remote sensing tab, two providers were supported in the final version, the BDC and MAPAQUALI projects. Both providers were selected due to their relevance to the BONDS project as they provide Amazon related data. These providers are being developed in the same infrastructure as they share common design patterns. We used *pystac-client* package to create a connection to the providers to browse their STAC and discover images and data cubes. The images are delivered by a Geoserver as COG, allowing GDAL to dynamically load image subsets, thus reducing the charge on the local QGIS instance.

Search can be refined through filters by collection, date, and spatial bounding box. Also, the user has the option to use the bounding box of a previously loaded QGIS layer (Figure 3b). Once the users select an item, its thumbnail (preview) and assets (bands) will be displayed and ready to be loaded to QGIS or downloaded. For BDC it is mandatory to have an API key to access the files. The key is available for registered users and can be generated at BDC web portal under the user's menu. Standardized access and discovery of data, through STAC allows to extend the plugin to reach other providers. GDAL's dynamic load proved to be an important feature to reduce the data load at the QGIS application. However, the size of remote sensing data is still a problem for data analysis and information extraction especially when you need all the data within the GIS (e.g. time series) [Ferreira et al. 2020].

3.3. BONDS field data

BONDS' field data are diverse and complex due to their data structures and the pre-processing steps needed to make data ready for use. With this heterogeneity, having a centralized common structure between datasets is very difficult. In this manner, the use of LabISA's integrated water quality database helps us to better understand and manage field data.

LabISA's tabular structure was composed of two main tables with field works and sample station information and several tables with water quality parameters linked to the main tables. QT and QGIS tools allow users to create a connection and perform

queries within PostgreSQL given a valid username and password. In the GUI users have the possibility to filter the data by the desired quality parameter and also by fieldwork (time-related) (Figure 3c). After selection, the filtered data is displayed in a tabular structure and can be loaded to QGIS as a georeferenced point layer or downloaded as CSV. Using LabISA's PostgreSQL database with QGIS and QT tools was successful, however, publishing this data as OGC standards presents a more standardized and reliable way to access this field data in the future.

3.4. Health COVID-19

For COVID-19 health data, two providers and 6 products were implemented in the plugin. First, the global COVID-19 data bulletins from CRC, products are: confirmed, recovered, and death cases and also vaccination status (Figure 3d). Regarding Brazil's data, the OpendataSUS daily bulletins and the historical resumed panels were selected. These summarized data are already screened and have been available since the beginning of the pandemic.

3.5. Utils

During the plugin development, some tools were identified and added to the Utils tab to improve user's experience with the QGIS environment. This tab has currently three functionalities implemented: a) Basemaps loading; b) CSV to point layer converter; c) Brazil Open Street Map (OSM) data download. The Basemaps used in the BONDS plugin is distributed as a Tiled Web Map (TWM) which supports the loading of several tiles in parts, making it seemingly fast. Also, a simple CSV to layer converter was added to help users load local data. The OSM data is a bundle of several geospatial data from Brazil that is maintained and distributed by Geofabrik ⁴.

4. Conclusion

In this paper we presented a QGIS plugin for discovering, browsing and acquiring (download) datasets from several data providers with different sharing technologies. The use of QT Designer within QGIS application was helpful to design the BONDS plugin GUI as a non-programming view. It is important to note that the GDAL within QGIS was essential to the project development as it presents a reliable and widely used Toolkit for geospatial handling. Also, the QGIS Python API presented a useful set of tools for data import, export, and management within the QGIS application.

The use of geoservices and cloud databases with standardized infrastructures and protocols has proven to be a great way to deal with complex geospatial data. These services often not only provide an easy access to data but also several server-side tools to aid users. Also, in modern technologies, the possibility of requesting only the necessary data can greatly improve the user's experience. Although geoservices offer great tools it is important to note that there are also some drawbacks. The first is a dependency on good internet connections from both servers and users. Second, the users are prone to changes in the service standards or data distribution policies. Lastly, as those technologies are still maturing their support and documentation can be sometimes absent.

⁴<https://www.geofabrik.de/>

We believe that the use of standardized web services and geoservices are already the main option to deal with data analysis. However, at the Client layer, the necessity to download this data locally in the GIS presents a bottleneck for large datasets or time series. On this matter, some approaches are exploring the concepts of server-side processing, such as the WPS and WCPS. Following the moving code paradigm, these approaches aim to bring new levels of abstraction for users to perform data analysis directly on the server [Müller et al. 2010, Müller 2016]. However, at the current state of data availability for heterogeneous datasets and providers, server-side approaches are still not feasible. As technologies and standards for geoservices mature, their uncertainties are constantly being reduced, and new solutions are proposed. At the current state, the BONDS plugin was assertive in centralizing some of the available services and providing tools for BONDS researchers to discover, access and download the datasets.

5. Acknowledgements

The authors would like to thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001 for funding his Ph.D scholarship. Also, thanks to CAMPUS FRANCE Brazil and SABERES project (BNP Paribas) for co-funding this research. Thanks to LabISA and BONDS (Biodiversa - Belmont Forum) for providing access to data.

References

- Baumann, P. (2010). The ogc web coverage processing service (wcps) standard. *Geoinformatica*, 14(4):447–479.
- BRASIL (2022a). Campanha nacional de vacinação contra covid-19 [national vaccination program against covid-19]. <https://opendatasus.saude.gov.br/dataset/covid-19-vacinacao/>.
- BRASIL (2022b). Opendatasus. <https://opendatasus.saude.gov.br/>.
- CONCAR (2010). Plano de ação para a implantação da inde.
- de Lucia Lobo, F., Barbosa, C. C. F., de Carvalho, L. S., Menino, F., Carlos, E. M. L. d. M., and Novo, E. F. d. S. (2017). Proposta de um sistema de monitoramento integrado da qualidade das águas continentais com imagens de satélites: Mapaquali.
- Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534.
- Ferreira, K. R., Queiroz, G. R., Vinhas, L., Marujo, R. F., Simoes, R. E., Picoli, M. C., Camara, G., Cartaxo, R., Gomes, V. C., Santos, L. A., et al. (2020). Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24):4033.
- Giuliani, G., Dubois, A., and Lacroix, P. (2013). Testing ogc web feature and coverage service performance: Towards efficient delivery of geospatial data. *Journal of Spatial Information Science*, 1(7):1–23.
- Gottschalk, K., Graham, S., Kreger, H., and Snell, J. (2002). Introduction to web services architecture. *IBM systems Journal*, 41(2):170–177.

- INDE (2022). Catálogo de geoserviços. <https://inde.gov.br/CatalogoGeoservicos>.
- Lawhead, J. (2017). *QGIS Python Programming Cookbook*. Packt Publishing Ltd.
- Lee, J.-G. and Kang, M. (2015). Geospatial big data: challenges and opportunities. *Big Data Research*, 2(2):74–81.
- Lemmens, R. (2006). Semantic interoperability of distributed geo-services. ITC.
- Li, D. (2008). On generalised and specialised spatial information grids: are geo-services ready? *International Journal of Digital Earth*, 1(4):315–325.
- Li, D., Shao, Z., Zhu, X., and Zhu, Y. (2004). From digital map to spatial information multi-grid. In *IGARSS 2004. 2004 IEEE International Geoscience and Remote Sensing Symposium*, volume 5, pages 2933–2936. IEEE.
- Maguire, D. J. and Longley, P. A. (2005). The emergence of geoportals and their role in spatial data infrastructures. *Computers, environment and urban systems*, 29(1):3–14.
- Michaelis, C. D. and Ames, D. P. (2012). Considerations for implementing ogc wms and wfs specifications in a desktop gis.
- Momjian, B. (2001). *PostgreSQL: introduction and concepts*, volume 192. Addison-Wesley New York.
- Müller, M. (2016). Service-oriented geoprocessing in spatial data infrastructures.
- Müller, M., Bernard, L., and Brauner, J. (2010). Moving code in spatial data infrastructures—web service based deployment of geoprocessing algorithms. *Transactions in GIS*, 14:101–118.
- Nascimento, H. L., de Souza Baptista, C., de Andrade, F. G., and Santos, L. C. (2020). Towards a resilient spatial data infrastructure. In *GeoInfo*, pages 34–45.
- OGC (2022). Open geospatial consortium. <https://www.ogc.org/about>.
- Raju, P. (2006). Fundamentals of geographical information system. *Satellite Remote Sensing and GIS Applications in Agricultural Meteorology*, page 103.
- Santos, A. F. D. (2005). The policy for commercializing cbers data. In *56th International Astronautical Congress of the International Astronautical Federation, the International Academy of Astronautics, and the International Institute of Space Law*, pages E6–1.
- Silva, P. N. and Pinheiro, M. M. K. (2019). Execução da política de dados abertos no brasil. *Tendências da Pesquisa Brasileira em Ciência da Informação*, 12(2).
- Stopa, S. R., Malta, D. C., Monteiro, C. N., Szwarcwald, C. L., Goldbaum, M., and Cesar, C. L. G. (2017). Use of and access to health services in brazil, 2013 national health survey. *Revista de saude publica*, 51.
- Tole, A. A. et al. (2013). Big data challenges. *Database systems journal*, 4(3):31–40.

Spatiotemporal Representation of the Territorial Evolution of Brazil

Fernanda de Oliveira Ramalho¹, Clodoveu A. Davis Jr.¹

¹Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brasil

{fernanda.ramalho, clodoveu}@dcc.ufmg.br

Abstract. *This paper presents the modeling of a geographic database that includes temporal data on the evolution of territorial boundaries. Data on the changes that occurred in the Brazilian territory between the years of 1872 and 1991 are modified in order to implement the necessary structures for manipulation of temporal data. This work's purpose is to achieve a single structure for analysis, comparison and visualization of spatiotemporal data. The result of the modeling is presented, along with an interaction in which it is possible to retrieve geographic data from a range of years.*

Keywords: *temporal data, spatiotemporal data, geoprocessing, GIS, Brazilian territory, database modeling*

1. Introduction

Brazil is currently composed of 26 states plus the Federal District and 5,570 municipalities, according to IBGE¹. Over the years, the Brazilian territory has changed numerous times, either by changes in the country's borders, or by modification of internal subdivisions, such as states (or provinces) and municipalities. These changes include dismemberment or union of municipalities, state divisions, and changes in subdivision names.

Municipalities can emerge from the territory of other municipalities, in a process that subdivides the territory and introduces a new federal unit. Likewise, municipal or state territories may be combined, or merged into one another. Such situations are described in an IBGE document², which also contains information on name changes, some of which are due to a reform of the Portuguese spelling in the early twentieth century.

There is a challenge in the modeling of this historical process in a spatial database. In fact, due to the modifications in the shape of subdivisions over time, spatiotemporal representation is required, so that the territorial configuration of the country can be retrieved at any point in time. This requires modeling temporal aspects of the data that are associated to geographic representations, and the definition of a set of integrity constraints that ensure that, for any reference date, there is a complete and correct territorial configuration. The implementation of the ensuing database would enable historians and students to gain a better view of the evolution of Brazil's territorial division.

According to [Dias et al. 2005], spatiotemporal modeling is one of the biggest challenges of geoinformation. Such modeling must be capable of representing changes

¹<https://www.ibge.gov.br/> - Instituto Brasileiro de Geografia e Estatística

²<https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/15771-evolucao-da-divisao-territorial-do-brasil.html>

both in space and in time. Resulting schemas are based on two main features: the choice of space and time representation concepts, and the construction of appropriate computational representations.

This work aims to present a spatiotemporal database schema that represents the changes that have happened in the Brazilian territory over the years, ensuring the integrity of the data. Since no complete spatiotemporal database management system exists today, the schema is implemented using a geographic database management system, namely PostGIS, with the creation of special attributes that allow selecting the set of valid territorial subdivisions at any given time. This allows analyzing and comparing the changes over the years, and for that purpose we present a visualization tool attached to the resulting database.

2. Related Works

Concerning spatiotemporal databases, there is no consensus on how to represent the temporal informations. As discussed by [Ferreira et al. 2015, Santos et al. 2016], there are many proposals of conceptual models to represent spatiotemporal data in GIS, but most existing technologies are specific for certain application domain.

[Gregory 2002] discusses that many European countries and regions created or were planing to create geographic systems in order to understand the changes of their administrative boundaries over time. The author examines different approaches to creating temporal GIS, the difficulties and the advantages of having spatiotemporal databases, then compares them with the solutions and methods used by the European countries to add a temporal dimension to their GIS data model. The approaches used involve digitise boundaries at key dates and creating a space-time composite structure, both similar to [Langran 1989]’s proposed models to a spatiotemporal GIS, and date-stamping individual features, based on [Vrana 1989]’s approach.

[Plumejeaud et al. 2011] presents a data model that manage changes in territorial organization, regarding territorial units split, merge and reorganization, causing changes in their limits. The model enables comparison of units and its changes over time. It is based on the identity of geographical units and their genealogy.

In this work, we propose a spatiotemporal model to represent the evolution of the Brazilian territory, focusing on a conceptual and physical modeling of a geographic database, incorporating a time dimension, based on [Meinerz 2005]’s Temporal OMT-G.

3. Material and Methods

3.1. Data

The previously mentioned document on the evolution of the state and municipal divisions of Brazil, produced by IBGE, describes the limits of the territories between 1872 and 1991. IBGE shows changes that occurred during these years, including population numbers and different spellings and names, and provides static maps of the subdivisions at each Census year. These maps were revised using modern techniques, so that borders that were preserved are consistent throughout them.

The data also contains information on the dismemberment of municipalities, including the source of the new subdivision and the names of each subdivision at the end

of the process, presented as an alphanumeric table. Figure 1 shows the difference in the division of the municipalities in the Brazilian territory between the years of 1872 and 1991. It is possible to notice the changes of the national borders and the difference in the number of municipalities.

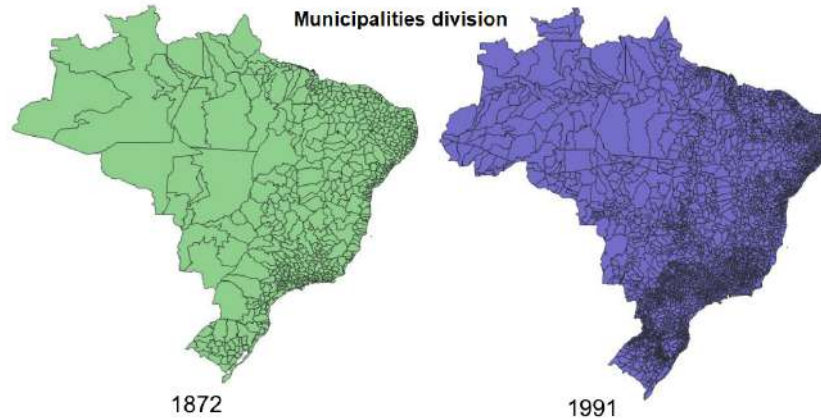


Figure 1. Brazilian territory in 1872 and 1991, divided into municipalities

3.2. Conceptual and Logical Modeling

The traditional models have two dimensions: data instances and their attributes. Each attribute has only one value, and if this value is changed, the previous one is lost. With temporal models, there is one more dimension, the temporal one, in which every new value is added to the database without removing the previous one [Edelweiss 1998].

There are two distinct time representations in a DBMS: valid time, which indicates the moment a change occurred, and transaction time, which is the moment when the change event was recorded in the database. A temporal database includes data on these moments, representing temporal information. For this work, as discussed later, valid time is more relevant, as it is used to indicate when a subdivision's boundaries changed.

There are two main classes to model the territorial subdivisions: "State" and "Municipality". Each class includes its boundaries, as its geographic representation alternative, plus the division instance's name and the year in which it has been created. For this model, it is necessary to have a DBMS that records the valid time. Thus, it is possible to access the data history, and the time can be associated with each value or instance.

The figures (Figure 2, Figure 3) show how the OMT-G modeling [Borges et al. 2001] for the State and Municipality elements would be done for a non-temporal geographic database. Figure 2 considers the two classes being represented by polygons and the aggregation relationship between them, showing that the State is composed by Municipalities – a schema that is independent of time. In Figure 3, the geographic part was separated, thus obtaining a non-geographical class for State and Municipality, linking each one to multiple geographic representations. Such alternative representations should then be used to represent evolution along time, using appropriate attributes that would indicate the period during which that subdivision remained valid.

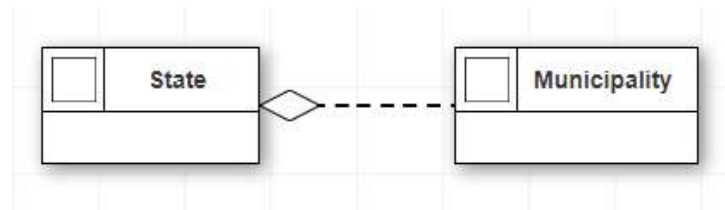


Figure 2. Modeling for a static geographic database

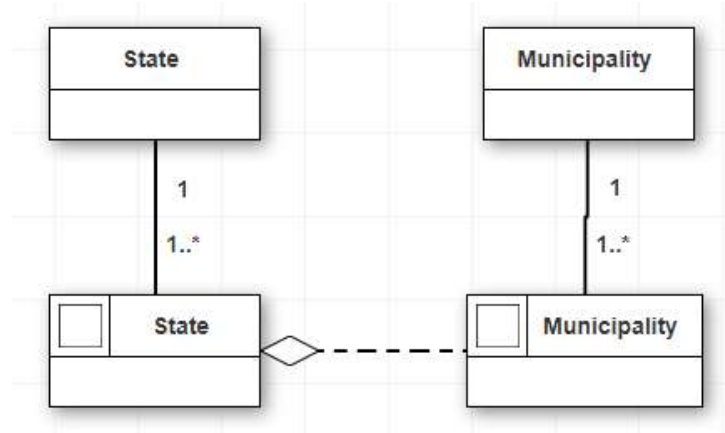


Figure 3. Modeling for a static geographic database (separated geography)

These two schemas, however, do not allow comparisons of previous states, nor do they allow for representing changes in attributes, such as municipality or state name. The separation made in Figure 3 simply enables the construction of a time series of geographic representations for State and Municipality.

To evolve from that limitation, we use elements of the Temporal OMT-G technique proposed by [Meinerz 2005]. This technique extends the OMT-G model to include primitives for the representation of spatiotemporal properties of geographical data. In Temporal OMT-G, stereotypes are used to indicate the model of time used in the schema. The technique adopts, among others, the transaction time and the valid time. For this work, the valid time was considered, which allows recording current, past and possibly future states of the classes, their representations and their attributes.

According to [Meinerz 2005], valid time defines Value Primitives, that are used to define new properties for elements of the model, and Constraints. Value Primitives refer to the valid time's lifecycle, which requires the time value to be entered by the user. They also refer to the Temporal Interval, that represents the duration of time records. Constraints refer to the length of time the element has existed in the database and require that the element only contains static and/or valid time properties; and that the values exist in the database only during their lifetime. Temporal OMT-G allows modeling the temporality of relationships, so that the association between instances of the involved classes can vary along time.

Figure 4 shows the result obtained by applying the temporal aspects to the conceptual model. Since Temporal OMT-G does not require all attributes to vary in time, we

decided to have both temporal and static attributes on the same object. The Valid Time was added to the State and Municipality attributes, in order to show that the attributes such as “Name” of the objects have the temporal property.

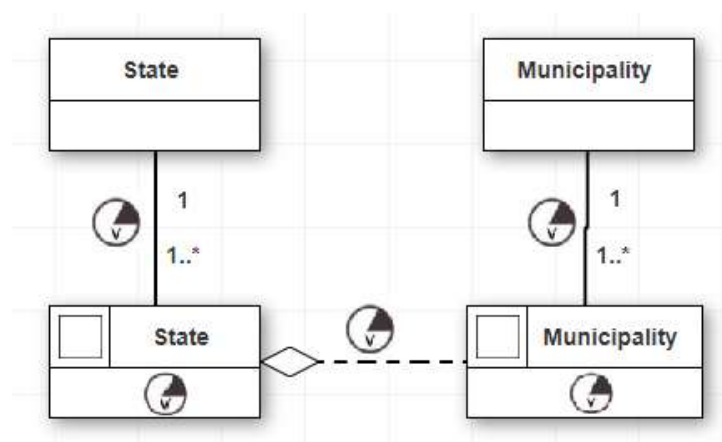


Figure 4. Conceptual modeling for a temporal geographic database

Between the conventional and spatial elements, temporal relationships with Valid Time were adopted. This makes it possible to maintain various instances of the relationship between objects over time. One example would be in the case of municipalities that move to another state.

The aggregation between the georeferenced elements is of Valid Time and shows that “State” can consist of many “Municipalities”. The temporal aspect is also important in this case to show that, even if the “Municipality” changes its geography and, for example, is no longer contained within “State”, its integrity is guaranteed.

3.3. Physical Modeling

3.3.1. Tools and Programs

The softwares used in this work were PostgreSQL³ to manage the database and the PostGIS⁴ extension, which allows the storage and analysis of GIS (Geographic Information System) objects in databases.

To handle the data, scripts were developed in Python and SQL. QGIS⁵ was used for the visualization, editing and analysis of georeferenced data, in addition to ArcGIS⁶, which also serves for the visualization and analysis of geographic information.

3.3.2. Data Treatments

From the conceptual modeling obtained, it was necessary to restructure the data so that valid time could be used in the final schema.

³<https://www.postgresql.org/>

⁴<https://postgis.net/>

⁵<https://qgis.org/>

⁶<https://www.esri.com/arcgis/>

Data provided by IBGE were separated into tables indicating the reference year for each object, and had the name of the municipality, municipality code and geometry as attributes. In order to implement the conceptual model, while making use of the idea of temporality, each one of the tables were extended to contain a start date and an end date to indicate the period through which the boundaries remained valid. Since the data is still separated by year, these dates are the same.

The start and end dates are related to any aspect of the municipality. If the geometry or name changes, the expiration date also follows those changes.

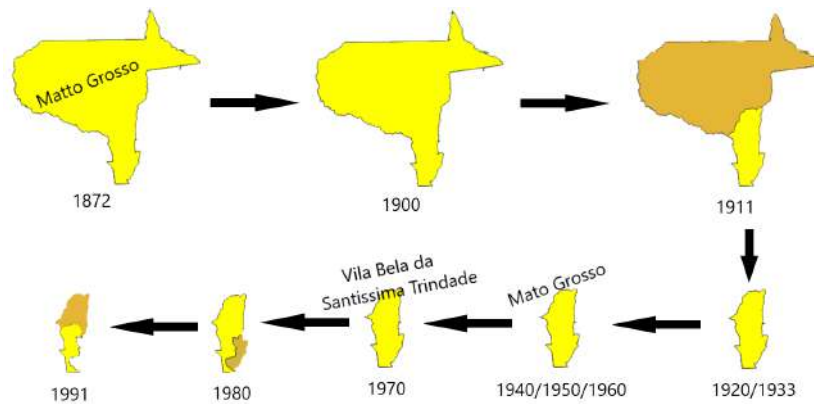


Figure 5. Municipality changes during the years. Same code, but different names and geometries

Figure 5 represents a municipality during the registered years. It has the same code, thus is treated as the same municipality across all the initial tables. But it also has geometry (between 1900 and 1911; 1970 and 1980; 1980 and 1991) and name (between 1933 and 1940; 1960 and 1970) changes. In the next sections we discuss how these changes were incorporated into the temporal modeling.

3.3.3. Geometric Aspects

Each municipality has a polygon associated with it, which represents its territory.

According to IBGE, the maps with the evolution of the municipalities were built from the Institute's data on municipal borders in 1991. Older versions of the municipal territories were built incrementally from 1991, back to 1872, since the most common modification is the splitting of a municipality in two, with the emancipation of a district or secondary urban nucleus. Thus, the municipal subdivisions were assembled from the most recent year to the oldest, through the union of polygons. This process is intended to avoid topological inaccuracies between different years, but IBGE warns, in the documentation of the process, that small issues persisted.

For this work, to put together a series of temporal versions of each municipality, it was necessary to compare the geometries in sequence along the tables. IBGE indicates that the codes of the municipalities are defined in such a way as to be unique; thus, when

a municipality is extinct, its code cannot be assigned to another. Likewise, municipalities that have changed states receive new codes.

A sequence of temporal versions of municipalities can then be put together, based on the IBGE code. If a sequential pair of versions has the same name and the same geometry, they are merged, using the initial valid date of the first version and the final valid date of the second. This new, merged, version is then compared to the next temporal version. If there are differences, the process stops, and the version is added to a results table (see next section). The divergent version is then used as the start of a new sequence of temporal versions, up until 1991.

In order to check whether the geometries of two versions actually coincide, PostGIS topology functions were used. The ST_Equals function was first used, checking if the geometries are spatially equal. However, as there are certain inaccuracies in the geometric data, the results obtained were incomplete.

To get around this situation, we compare the area of the difference between the two polygons with the area of one of them. Differences under a small threshold were disregarded. Table 1 shows some of the results of the comparisons of versions of municipalities between the years 1940 and 1950. It presents the name of the municipality in each year, its code and the result of the area difference.

Name (1940)	Name (1950)	Code	Area Difference (%)
Porto de Pedras	Porto de Pedras	2707404	0
Mineiros	Mineiros	5213103	1.76169066618888e-09
Santa Rita	Santa Rita	2513703	0.029405652021209

Table 1. Table with the calculated differences of the geometries of two municipalities

When the difference is zero, it means that geometries can be treated as being spatially equal. Differences like the one in “Mineiros” were considered to be irrelevant. The municipalities that had these results are those where the projection was slightly different, but still the same. When the difference is not as close to zero, as in the municipality of “Santa Rita” shown in the table, it means that the municipality may or may not have territorial differences. To make this distinction, experiments were carried out by varying a tolerance limit of the difference values.

The query below shows how the difference between the geometries of two years were calculated. The geometry type is “Multipolygon”, therefore it was possible to use PostGIS’s functions ST_SymDifference and ST_Area. Tables ‘a’ and ‘b’ are tables of municipalities of consecutive years, i.e. 1940 and 1950 respectively. The total difference is represented by the area of the difference between the two geometries divided by the area of the first geometry, resulting in a percentual value for each entry.

```
SELECT a.code, a.name, b.name,
       ST_Area(ST_SymDifference(a.geom,b.geom))/ST_Area(a.geom)
       as diff
FROM table_a as a join table_b as b
ON a.code = b.code
ORDER BY diff;
```

The threshold was obtained by observing the results of the query and analyzing the municipality areas, with the help of the visual tool in QGIS, and identifying from which percentual there were significant differences between them, indicating changes in the territory.

Table 2 presents an example of part of the query's result for the years of 1940 and 1950, assuming that the name is the same. After an analysis, it was observed that the borders of "Santa Rita" and the municipalities above it did not have visual changes, while "João Pessoa" and the municipalities below it had significative changes on them. Thus the threshold adopted for this example is approximately 3%, i.e. all the entries with differences less than this are considered to have had no geometry changes. This verification was made for each year change, and a new threshold was obtained each time.

code	name (1940/1950)	diff (%)
3099901	Distrito Federal	0.0143346706289836
3302007	Itaguaí	0.0234711420848194
2513703	Santa Rita	0.029405652021209
2507507	João Pessoa	0.0337054292980466
2708006	Santana do Ipanema	0.0645548440398625
4113205	Lapa	0.0750746375691199
3149903	Perdões	0.0769926575000797

Table 2. Result of the query containing the differences between the areas (1940 and 1950)

The geometric difference between A and B is different from the geometric difference between B and A. Thus, to avoid erroneous results, the symmetric difference between the geometries was used (ST_SymDifference).

From the comparisons of geometries, it is possible to visualize how the municipalities are incrementally created. Parts of the territory may become emancipated, others can be annexed or incorporated to an existing municipality, or municipalities can be merged, losing their individual characteristics.

3.3.4. Temporal Aspects

Municipalities that resulted from the process described in the previous section were stored in a single table. The table's structure is shown in the figure 6. In this structure, the table no longer holds the physical version of an OMT-G planar subdivision, but a set of them. Each polygon has a creation and an expiration date. These attributes indicate the time period during which that version (geometry, name, code) was valid. Since the national territory has changed along history, new versions of the state borders are then obtained from the union of municipality polygons, grouping them according to the first digits of the code. National borders can be obtained in the same way.

Figure 7 shows a sequence of changes that occurred during three censuses in the region where Belo Horizonte is today. Table 3 shows how these changes are encoded in the database table.

divisions	
id	int
code	float
name	varchar
geom	geometry(Multipolygon)
creation_date	int
expiration_date	int

Figure 6. Scheme considering the temporal aspect



Figure 7. Example of a sequence of changes over the years

As can be observed in figure 7, in 1872 there was only the municipality of Sabará. So the creation date reflects the situation in 1872. When analyzing the next year, 1900, it is observed that the municipality of Sabará was divided into three. As Sabará has a new geometry, an updated entry is created on the table, with a creation date of 1900. The expiration date is now 1991, since no other modifications have occurred in Sabará since this split.

From part of Sabará, a new municipality, Belo Horizonte, appeared in the 1900 data. Thus, we assume this to be its creation year. Between 1933 and 1940, the city's name changed to "Belo Horizonte", with the double letter removed by an orthographic reform of the Portuguese language. No modifications to the territory were observed. The original version then has an expiration date of 1940, and the new version is considered to be created in 1940. The same types of modifications occurred in Nova de Lima. First, two name changes, one between 1900 and 1911, and the other between 1920 and 1933; and then a second division, between 1940 and 1950, creating Raposos and Rio Acima.

Notice that changes are perceived when the version corresponding to a new Census is different from the one preceding it. Thus, a change must have occurred in the years between the two Census years. Establishing and recording the actual data of the change is left for future work, as this requires further data. The actual date for the founding of Belo Horizonte is December 12, 1897, and the orthographic reform took place in 1938.

As there were no changes in the municipality of Sabará from 1900 onwards, the expiration date is updated to the maximum value, 1991. The same goes for the most recent entries of Belo Horizonte, Nova Lima, Raposos and Rio Acima. Expiration dates

Name	Creation Date	Expiration Date
Sabar	1872	1900
Sabar	1900	1991
Bello Horizonte	1900	1940
Belo Horizonte	1940	1991
Nova de Lima	1900	1911
Villa Nova de Lima	1911	1933
Nova Lima	1933	1950
Nova Lima	1950	1991
Raposos	1950	1991
Rio Acima	1950	1991

Table 3. Table with examples of changes that happened over time

with a null value are used to indicate the current version of each record.

3.4. Final Physical Modeling

With the grouping of all data in a single structure, it is possible to have more than one entry for the same municipality. Therefore, the IBGE code can no longer be used as a primary key. The final version of the table then uses a composite key, comprising the code and the creation date.

`municipalities(code, name, geom, creation_date, expiration_date)`

4. Interaction

With the objective of comparing and visualizing, in a more concentrated way, the changes that occurred in the Brazilian territory, an interactive visualizer was developed using the ArcGIS application tools. Figure 8 shows the initial screen of the developed interaction.

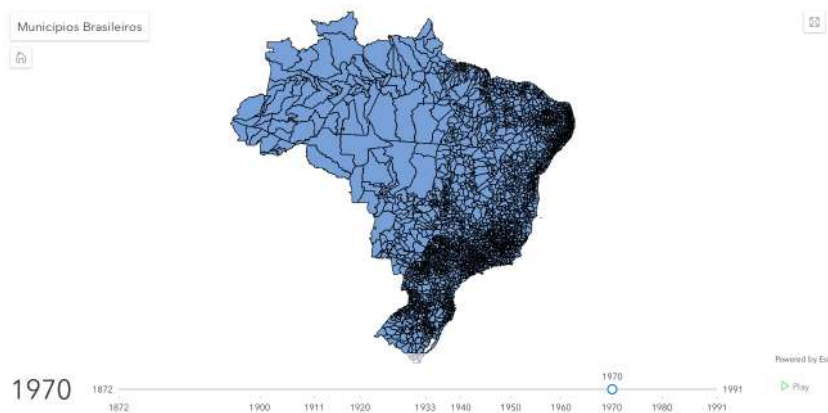


Figure 8. Screen of the interactive interface

It has a timeline with the years between 1872 and 1991, in which it is possible to select each available year and visualize the corresponding territorial configuration. It also has a *play* button that allows the user to observe the evolution of the territory as an animation.

The interaction can be found at:

<https://feramvalho.github.io/Municipios/MunBR/>

5. Conclusion and Future Works

The integration of geographic elements with temporal elements resulted in a concise structure that allows the maintenance of different versions of the territorial configurations, without unnecessary repetition. The insertion of creation and expiration dates facilitated the comparison and analysis of the municipalities over the years.

This structure allows the easy recovery of the history of each part of the territory, with the possibility of comparing them to other moments and directly analyzing the changes that have occurred. The implementation of the interaction was simplified because of the physical implementation, which takes into account the temporal aspects.

The final modeling and concepts can be used to facilitate analysis and representations that ensure data integrity, contemplating the spatial and temporal aspects of the data. All changes incorporated into a structure can benefit researches involving retrieval of different territorial situations, and also assist in studies of the Brazilian territory.

Since there are no structures today that present a concise representation of spatiotemporal data while allowing observation and tracking of its changes over time, this work, in addition to presenting an alternative for modeling territorial changes, also offers a starting point in modeling more generalized structures. The use of Temporal OMT-G including valid time was crucial to the development of the conceptual modeling, which was later transformed in the database's physical modeling.

To extend this work, we intend to add the modifications after 1991, and also improve the data to reflect the real creation dates of the municipalities.

For future work, the idea is to apply the concepts studied and the modeling made in larger structures, including other situations of territorial evolution, such as in rural areas. It is also intended to evolve the modeling and to improve the interaction, adding more options for analysis.

References

- Borges, K. and Davis, C. (2002). Modelagem de dados geográficos. *CÂMARA, G.; MONTEIRO, AM; DAVIS, C. Geoprocessamento: teorias e aplicações. 3v, 3.*
- Borges, K., Davis Jr, C., and Laender, A. (2001). Omt-g: An object-oriented data model for geographic applications. *GeoInformatica*, 5.
- Borges, K. A., Davis Jr, C. A., and Laender, A. H. (2005). Modelagem conceitual de dados geográficos. *CASANOVA, et. al. Banco de Dados Geográfico. MundoGEO: Curitiba*, pages 83–136.
- Correia, A. H., Piromal, R. A. S., QUEIROZ, G., and SOUZA, R. (2005). Modelagem de um banco de dados espaço-temporal para desmatamento e queimadas. *XII Simpósio Brasileiro de Sensoriamento Remoto, Goiânia.*
- Dias, T. d. L., Câmara, G., and Davis Jr, C. A. (2005). Modelos espaço-temporais. *MA Casanova, G Câmara, CA Davis Jr, L Vinhas, GR Queiroz.(Org. e Ed.). Bancos de Dados GeoGráficos. Editora Mundogeo, Curitiba, Cap, 4.*

- Edelweiss, N. (1998). Bancos de dados temporais: Teoria e prática.
- Ferreira, K. R., de Oliveira, A. G., Monteiro, A. M. V., and de Almeida, D. B. (2015). Temporal gis and spatiotemporal data sources. In *GeoInfo*.
- Filho, J. L., Iochpe, C., Hasenack, H., and Weber, E. J. (2001). Modelagem conceitual de banco de dados geográficos: o estudo de caso do projeto padct/ciamb.
- Gregory, I. (2002). Time-variant gis databases of changing historical administrative boundaries: A european comparison. *T. GIS*, 6:161–178.
- Langran, G. E. (1989). *Time in geographic information systems*. University of Washington.
- Lisboa Filho, J., Iochpe, C., and Garaffa, I. (1997). Modelos conceituais de dados para aplicações geográficas: uma experiência com um sig interinstitucional. In *Simpósio Brasileiro de Geoprocessamento*, volume 4.
- Meinerz, G. V. (2005). Omt-g temporal: Uma técnica de extensão do modelo omt-g para representar os aspectos temporais de dados geográficos. Master's thesis, Instituto Tecnológico de Aeronáutica.
- Oliveira, M. P. (2007). *Visualização de dados geográficos dirigida pelo Modelo Conceitual OMT-G. 2007. 106 p.* PhD thesis, Tese (Doutorado em Computação Aplicada)-Instituto Nacional de Pesquisas.
- Peuquet, D. (1999). Time in gis and geographical databases. In *Geographical information systems*, volume 1, pages 91–103.
- Plumejeaud, C., Mathian, H., Gensel, J., and Grasland, C. (2011). Spatio-temporal analysis of territorial changes from a multi-scale perspective. *International Journal of Geographical Information Science*, 25(10):1597–1612.
- Santos, L. A., Ferreira, K. R., de Queiroz, G. R., and Vinhas, L. (2016). Spatiotemporal data representation in r. In *GeoInfo*, pages 178–191.
- Vrana, R. (1989). Historical data as an explicit component of land information systems. *International Journal of Geographical Information Systems*, 3(1):33–49.

Territorial relations between deforestation and Rural Environmental Registry (CAR) in the Amazon biome using free software QGIS/PostgreSQL/PostGIS and Data Warehouse Structure

Mayra Abboudi Brasco¹, Carlos Alberto de Carvalho¹

¹Embrapa Territorial

Avenida Soldado Passarinho, 303, CEP 13070-115 - Campinas - SP

mayra.brasco@colaborador.embrapa.br, carlos-alberto.carvalho@embrapa.br

Abstract.

This work presents the use and integration of free software to organize a spatial database (part of a data warehouse) capable of performing spatial and numerical analyses on a set of spatial data from the Amazon biome (deforestation and Rural Environmental Registry - CAR). This work is part of a larger research project (Embrapa/Censipam) that addresses the use of strategic territorial intelligence to analyze deforestation processes in the Amazon biome. More specifically, this article deals with a part of the Data Warehouse construction that serves as the basis for the larger research project to work. The focus is on the spatial part of the data warehouse, proposing a free software structure to perform spatial analysis. As preliminary results, quantifications of deforestation by state and by municipalities are presented.

Keywords: Amazon biome; Rural Environmental Registry; Data Warehouse; Quantum GIS; deforestation.

Resumo.

Este trabalho apresenta o uso e integração de software livre para organizar um banco de dados espacial (parte de um data warehouse) capaz de realizar análises espaciais e numéricas sobre um conjunto de dados espaciais do bioma Amazônia (desmatamentos e o Cadastro Ambiental Rural - CAR). Este trabalho faz parte de um projeto de pesquisa maior (Embrapa/Censipam) que aborda o uso de inteligência territorial estratégica para analisar os processos de desmatamento no bioma Amazônia. Mais especificamente, este artigo lida com uma parte da construção do Data Warehouse que serve de base para o funcionamento do projeto de pesquisa maior. O enfoque está na parte espacial do data warehouse, propondo uma estrutura de software livre para realizar as análises espaciais. Como resultados preliminares, são apresentadas quantificações de desmatamentos por estado e por municípios.

Palavras-chave: Amazônia; Cadastro Ambiental Rural; Data Warehouse; Quantum GIS; desmatamento.

1. Introduction

This article is one of the results of the project “Strategic territorial intelligence applied to the identification, quantification and qualification of deforestation processes and support and guidance of public and private policies in the Amazon biome”, a partnership between Embrapa Territorial and Centro Gestor e Operacional do Sistema de Proteção da Amazônia (Censipam). It seeks to broaden the analysis of deforestation processes in the Amazon biome through a detailed study of territorial and temporal bases, of deforestation patterns, such as the number of deforested polygons and their spatial and temporal evolution, concentration by state, microregion, legal attribution, proximity to

urban centers, logistics infrastructure, etc., and their territorial distribution according to the legal allocation of land.

The database for the project was built using the Kimball method and the scripts generated using ETL (Extract, Transform and Load) methods. The creation and structuring of the Data Warehouse were addressed in a previous work (DEUS; ALMEIDA; CARVALHO, 2021). In this paper, we will continue the analysis that supports the Data Warehouse created.

The present work consists of describing both the operational process of accessing the database created using the Data Warehouse concept, built on PostgreSQL/PostGIS and Quantum GIS, all free software, and the analysis of deforestation scenario between 2018 and 2020 in rural properties registered in the Rural Environmental Registry (CAR in the Portuguese acronym) (BRASIL, 2012). However, in this analysis the legal reserve areas within the CAR's polygons and the permanent protection areas around the rivers are not considered.

2. Free software: Quantum GIS (QGIS) and PostgreSQL/PostGIS

The use of free software and the movements for their democratization began a long time ago. In 1983, Richard Stallman created a movement in opposition to the Unix world, which created Linux operating systems, called GNU¹.

QGIS is open source licensed under the GNU General Public License and is a project of the Open Source Geospatial Foundation (OSGeo). Given this important feature, QGIS is constantly being updated, receiving contributions from communities through bug reports, plugins (or add-ons) that are created by software scholars and made available for free through the system's repositories.

PostgreSQL is an open-source, widely used relational database. For compatibility with other systems, the most stable version, 13.5, was used. PostGIS is the spatial extension to the PostgreSQL relational database for working with geospatial fields and functions. PostGIS - Spatial and Geographic Objects for SQL is a spatial extension for the PostgreSQL free database, the result of the same project as QGIS, OSGeo. It can be installed on any operating system, as well as QGIS, and there is an interface and direct connection between these software (Figure 1).

Free software from the OSGeo project, QGIS and PostGIS are tools that have a strong connection and are in constant evolution and adaptability. However, there is still resistance regarding their use for several reasons.

¹ At the time, Unix did not give freedom to its users. Currently, we see many versions of Linux that are more restrictive, with better interfaces and also those that allow users to access their source code at any time and make changes.

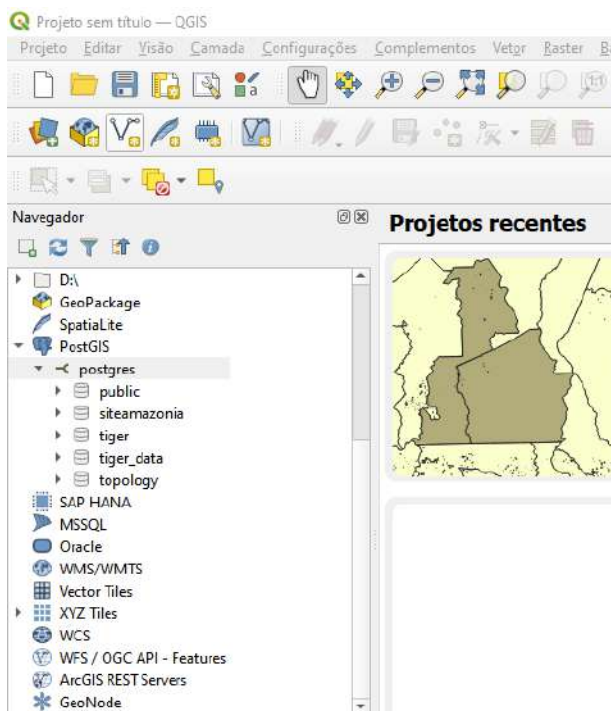


Figure 1. Screenshot of QGIS showing its direct connection with PostGIS.

Given the context of the project, we developed a method using these technologies. It gives anyone who wants to use the project's data the opportunity to analyze deforestation using free software. Therefore, with the continuity and the dissemination of free software methods, we aim to identify and quantify the numbers and dimension of deforestation that occurs annually in the Amazon biome, its states and municipalities, with specific analysis in areas registered in CAR.

3. Material and methods

This research is based on the Amazon biome and a GIS will be structured with it. This system will be the integration of software, tools and techniques with collections of numerical, iconographic and cartographic data, in GIS, supported by spatial databases.

The equipment used in this project have high storage capacity and performance: 5 HP Intel Core i7 3.4GHz Stations with 16GB of RAM, 500GB SSD HD and 2TB HD; 1 SY480Gen10 Blade Server with 10-core Intel Xeon Silver processors, 192 GB of RAM and 2 600GB HDDs and a large-capacity data storage system Storage HP MSA-2050 with a capacity of 192TB.

The software used in the integration and interconnection with the Data Warehouse structure of the project were Quantum GIS 3.24.1 Tisler, PostgreSQL 13.7 and PostGIS 3.1.

The data warehouse was structured using PostgreSQL and made available on an exclusive workstation with more computational resources and intended for the approval of the applications developed in the project. The hardware used has the following settings:

- Manufacturer: Lenovo;
- Brand: Thinkstation;
- Model: P520;
- Intel Xeon W-2255 Processor;
- Memory 64GB DDR4 ECC 2666MHz (4x 16GB);
- NVIDIA Quadro P2000 graphics card;
- 2TB 7200 RPM HD;
- SSD 512GB;
- Windows 10 Pro for Workstations 64 Operating System.

Windows 10 operating system contains the Windows Subsystem for Linux or WSL (Windows Subsystem for Linux) application [15], which “allows developers to run a GNU/Linux environment, including most command-line tools, utilities and applications, directly in Windows, without modification and without the overhead of a traditional virtual machine or dualboot installation”. Thus, we were able to migrate the Data Warehouse using the same Linux operating system where it was developed.

The limits of municipalities and states were obtained from the Brazilian Institute of Geography and Statistics’ (IBGE) territorial divisions as a shapefile (IBGE, 2014). The limits of the Amazon biome are the official ones of the Ministry of the Environment, cut from the shapefile of Brazilian biomes. The geocoded boundaries of Conservation Units, Indigenous Lands, Military Areas, Agrarian Settlements and *Quilombola* Communities were obtained from the National Spatial Data Infrastructure (INDE, 2022), digitally cut and structured using projection and nomenclature adjustments in the SIRGAS 2000 reference system (IBGE, 2010), and the projection adopted by IBGE (IBGE, 2014) for area calculations.

The geocoded data on rural properties in CAR will be obtained from the Brazilian Forest Service (SFB). From the National Register of Addresses for Statistical Purposes (IBGE, 2020), spreadsheets with the coordinates of agricultural establishments surveyed in the 2017 Agricultural Census will be obtained for each state and the Federation.

Deforestation by clear cutting in the Amazon biome has been monitored since 1988 by the Project for Monitoring Deforestation in the Legal Amazon by Satellite (PRODES), of the National Institute for Space Research (INPE) (INPE, 2022). Annual rates are estimated based on the deforestation increments identified in LANDSAT satellite images (20 to 30 meters of spatial resolution and revisited every 16 days). The minimum area mapped by PRODES is 6.25 hectares.

The CAR database used was the SICAR 2021 version. This database underwent a process of selection of polygons, and those contained in areas of integral conservation units, indigenous lands and military areas were removed. Then, the rural properties registered in the SICAR 2021 were intersected with data on deforestation between 2018 and 2020.

3.1. Database connection

As described at the beginning of the report, QGIS and PostgreSQL/PostGIS have an easy integration, because they are from the same developers. Therefore, their connection is easy and does not require any specific configuration.

To import a layer in QGIS, after fixing geometries and defining projection, it is necessary to click on the upper tab “Database” and import a layer using the “DB Manager”, then check the “primary key”, “geometry column”, “source SRID”, “target SRID” and “create spatial indexes” options. When importing it is necessary to define projection. For this database created using the Data Warehouse, we registered the IGBE_South_Albers_Equal_Area_Conic projection with the SRID code 1054, which is the IBGE's standard projection for calculating areas. Thus, when importing the layer, it is necessary to reaffirm to the database which projection is adopted using a command in the SQL language (Figure 2):

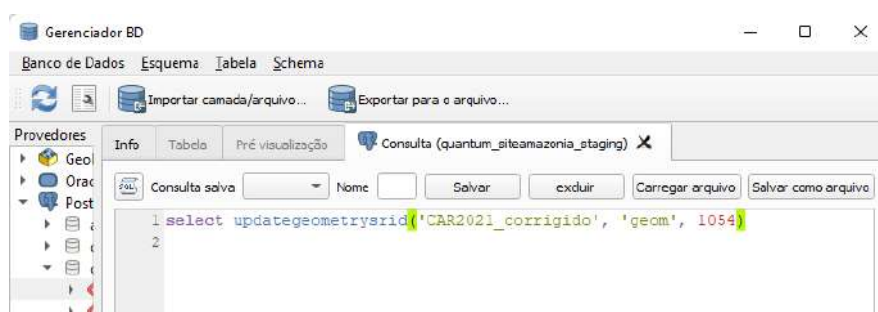


Figure 2. SQL command window in QGIS DB Manager.

Once the database is connected, it is possible to work on the layers from any computer that has access to QGIS, PostGIS and connection to the Data Warehouse built. This tool made it possible for team members of the project to connect and work remotely.

For the analyses depicted in this article we used QGIS tools to work on the data, more precisely the "dissolve" and "clip" tools. Then, using the "basic statistics" tool, we calculated the areas and quantities of polygons from the intersection of CAR 2021 data and deforestation polygons between 2018 and 2020. These were transferred to Microsoft Excel, where we made the tables and graphs presented below.

4. Analysis

Crossing the deforestation areas within CAR areas, their relationship with the total deforested areas and the numbers of deforestation polygons that coincide with the polygons of SICAR 2021, we observed a direct relation between particular rural properties and deforestation (Figure 3).

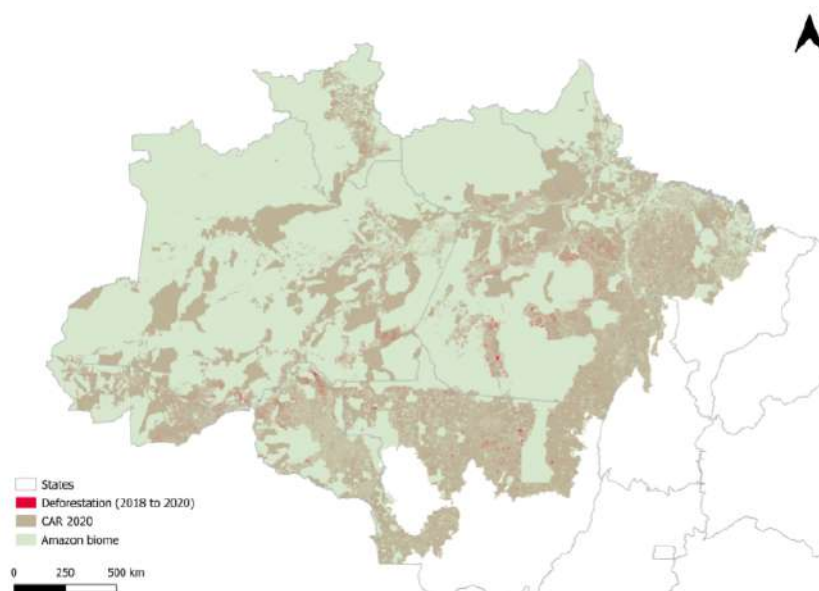


Figure 3. Amazon biome with CAR 2021 and deforestation (2018 to 2020)

Table 1 shows the general results: for each year, the total area deforested in hectares and the number of deforestation polygons, as well as the total area occupied by CAR 2021 and the number of rural properties and, finally, the deforested area that coincides with the area of rural properties, the percentage of deforestation in these areas that corresponds to the total area deforested in the year and the number of corresponding polygons.

Table 1. General results of the analysis of deforestation and SICAR 2021 (between 2018 and 2020)

Year	total area deforested in hectares	number of deforestation polygons	total area occupied by CAR 2021	number of rural properties - CAR 2021	deforested area that coincides with the area of rural properties - CAR 2021	percentage of how much deforestation of these areas occupied by CAR corresponds to the total area deforested	number of corresponding polygons that coincides with the area of rural properties - CAR 2021
2020	1.031.851	41.129	156.886.040	131.616	691.048	66,97%	6.646
2019	1.066.581	41.526	156.886.040	131.616	727.541	68,21%	6.750
2018	703.710	32.661	156.886.040	131.616	481.334	68,40%	6.011

We noticed that there was an increase of more than 300 thousand hectares deforested from 2018 to 2019, but a softening of the advance of deforestation between 2019 and 2020. However, there are still more than 1 million hectares deforested in each of these years.

When correlating the deforested areas with the areas occupied by rural properties, we also see that the amount of deforestation decreased within these areas in the period: in 2018, 68.4% of deforestation occurred within CAR areas; in 2019, 68.2%; and in 2020, 66.97%. Therefore, we see that most deforestation is carried out in these areas, but the increase in deforestation is inside and outside the CAR areas, as the increase is not proportional.

When we started to analyze the states, we noticed discrepancies between the deforestation numbers. In Figure 4, we see that some states have many deforested hectares, while others do not appear on the graph.

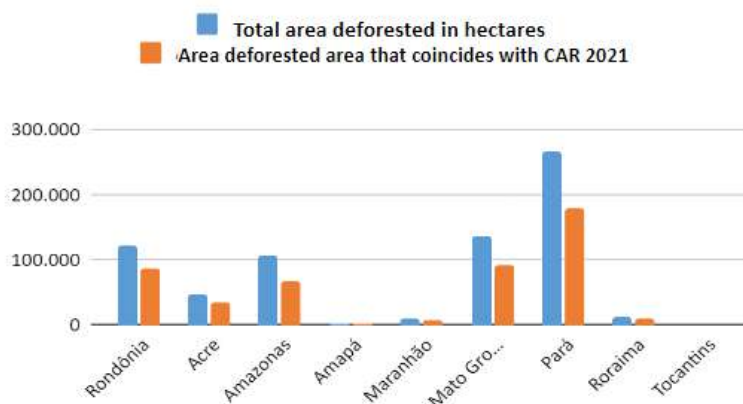


Figure 4. Graph - Deforestation that occurred in 2018 and its incidence within the rural properties of the 2021 CAR by states

The state of Pará had more of its area deforested in 2018, followed by Mato Grosso, Rondônia and Amazonas. Tocantins had less than a thousand hectares deforested in the year and does not appear in the graph. Most of the deforestation in the states occur within CAR areas. However, in some states, we have noticed that almost all deforestation occurs in rural real estate areas, such as Maranhão and Roraima.

When we analyze the 2019 data, presented in Figure 5, we see that Tocantins remains unimpressive, while Pará, Mato Grosso, Amazonas and Rondônia continue to be the states with the highest deforestation rates. But this time, Amazonas had a more deforested area than Rondônia. Most of the deforested areas are inside SICAR 2021, but we are beginning to notice a discrepancy in deforestation outside these areas in Pará.

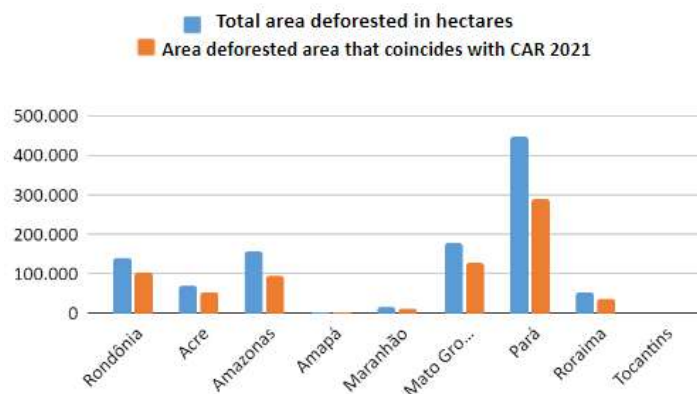


Figure 5. Graph - Percentage of deforestation carried out within the 2021 CAR areas in relation to the total deforestation carried out in the states in 2019

In 2020 (Figure 6), we notice few differences in these numbers. However, we see that the pattern of deforested areas remains the same, except for Maranhão, which starts to appear more in the graph for having a few thousand hectares more than in the other two years.

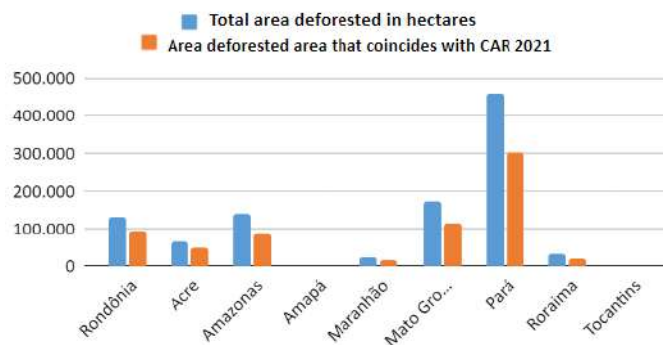


Figure 6. Graph - Percentage of deforestation carried out within the SICAR 2021 areas in relation to the total deforestation carried out in the states in 2020

When we start to analyze the municipalities, some other dimensions stand out. In Figure 7, there is a graph showing ten municipalities that deforested the most in 2018, but the percentages of deforestation that occurred within CAR areas are shown. Therefore, the calculation of the ratio between deforestation within SICAR 2021 and total deforestation produced the comparison in percentage.

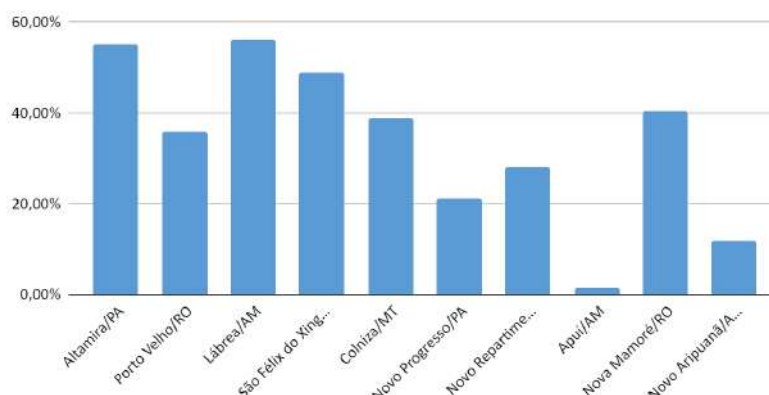


Figure 7. Graph - Ten municipalities with the highest deforestation rates in 2018 and how much of it (percentage) occurred in SICAR 2021 areas

Leaving the general analysis (states) and entering the specific analysis (municipalities), it stands out that not all municipalities have most of their deforestation taking place within rural properties. By studying the municipalities that deforest the most, case by case, we may think about local public policies.

We see that most of the deforestation in Porto Velho/RO, Colniza/MT, Novo Progresso/PA, Novo Repartimento/PA, Apuí/AM, Nova Mamoré/RO and Novo Aripuanã/AM occur in areas not registered in the Brazilian Rural Environmental Registry. A first analysis indicates the need for a focus on the inspection of areas outside the rural world. The available data could be verified with other data sources (limits of protected areas and indigenous lands).

In Figure 8, we see that the municipalities that stand out the most remain almost the same, with the exception of Pacajá and Portel in Pará. We see that the only city that had most of its deforestation within the CAR areas was Lábrea in Amazonas.

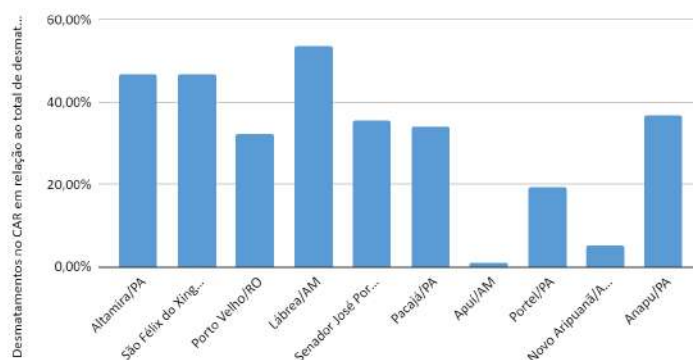


Figure 8. Graph - Ten municipalities with the highest rates (in 2019) of deforestation and how much of it occurred in SICAR 2021 areas

When looking at the 2020 numbers (Figure 9), we notice an increase in deforestation percentage in these areas, in the municipalities of São Félix do Xingu, Altamira and Lábrea, and the emergence of municipalities that were not yet on the list, such as: Novo Progresso and Itaituba, both in Pará. Most deforestation took place in CAR areas, with the exception of Porto Velho, Apuí, Pacajá and Portel. We perceive the need for greater control and inspection in these municipalities, in order to identify where the deforested areas are and make appropriate decisions to avoid illegal and unauthorized deforestation.

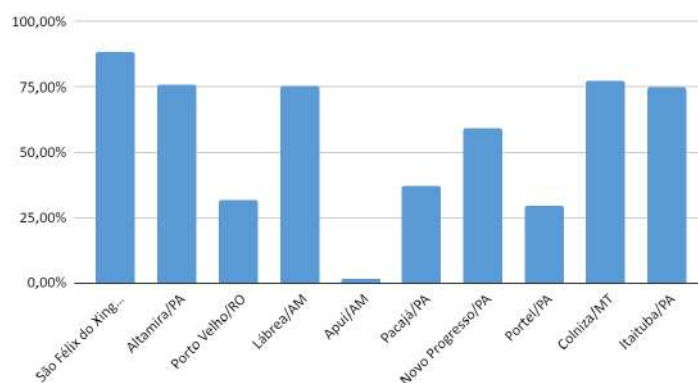


Figure 9. Graph - Ten municipalities with the highest rates (in 2020) of deforestation and how much of these occurred in SICAR 2021 areas

When spatializing the municipalities and deforestation polygons in SICAR 2021 (Figure 10), we see that they remain in the same range in the “arc of deforestation”. We also noticed an increase in polygons that coincide with the SICAR 2021 areas.

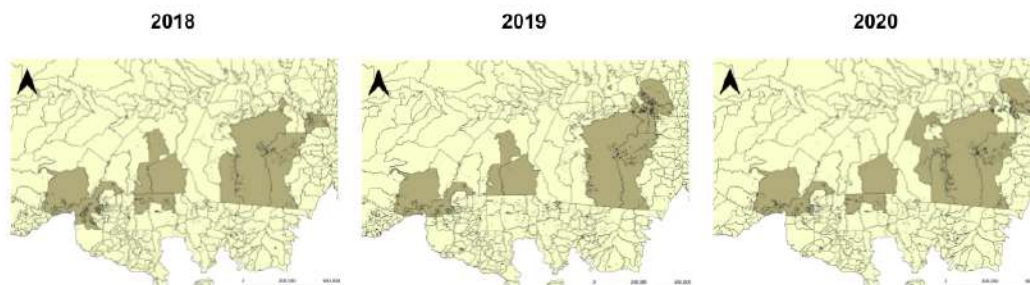


Figure 10. Map - The ten most deforested municipalities in the Amazon biome in 2018, 2019 and 2020 and deforestation polygons that coincide with the SICAR 2021 areas

This data was generated in this short period of research and approach to the theme, and there is a need to deepen concepts, analyses and other tools to deal with this complexity.

6. Conclusions

The use of free software in the project was an efficient and low cost solution. The possibility of creating a Data Warehouse according to the needs of the project and the generated data, its connection to Quantum GIS using the PostGIS extension and the great interoperability and ease of connection and processing of data offers what is necessary for the project to continue. With this structure in place, we were able to analyze deforestation in CAR areas over the temporal and spatial clipping described.

Approaching this research between deforestation and CAR using a territorial cutout detailed by municipalities opens a possibility of discussing public policies aimed directly at city managers, because it is where deforestation events actually occur. It also makes it possible to highlight cuts at the state level, enabling joint actions by different levels of public administration. By highlighting municipalities that concentrate deforestation and qualifying them in terms of their distribution in the rural world as portrayed by CAR, action may be taken to combat illegal deforestation.

In addition to the results, performing data processing and analysis using free software is also important, because when we work with information of public use, it is essential to disseminate results and methods that may be used and replicated, in order to foster the evolution of scientific knowledge and the elaboration of public policies by public institutions or companies.

References

BRASIL (2012). Lei n. 12.651, de 25 de maio de 2012. Dispõe sobre a proteção da vegetação nativa; altera as Leis n°s 6.938, de 31 de agosto de 1981, 9.393, de 19 de dezembro de 1996, e 11.428, de 22 de dezembro de 2006; revoga as Leis n°s 4.771, de 15 de setembro de 1965, e 7.754, de 14 de abril de 1989, e a Medida Provisória n° 2.166-67, de 24 de agosto de 2001; e dá outras providências.
http://www.planalto.gov.br/ccivil_03/_ato2011-2014/2012/lei/l12651.htm, August 10th, 2022.

DE DEUS, R.; ALMEIDA, M.; CARVALHO, C. (2021) Estruturação de um Data Warehouse para repartição territorial de desmatamentos no bioma Amazônia e o Cadastro Ambiental Rural. *In: CONGRESSO BRASILEIRO DE AGROINFORMÁTICA (SBIAGRO)*, 13. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 116-125.

GNU (2021). What is GNU?. <https://www.gnu.org/>, August 10th, 2022.

IBGE (2014). “Estrutura Territorial”. <https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial.html>, August.

IBGE (2010). “Projeto SIRGAS: Sistema de Referência Geocêntrico para as Américas”. <https://www.ibge.gov.br/geociencias/informacoes-sobre-posicionamento-geodesico/sirgas/16691-projeto-mudanca-do-referencial-geodesico-pmrg.html?=&t=o-que-e>, August.

IBGE. Cadastro Nacional de Endereços para Fins Estatísticos – CNEFE. Rio de Janeiro: IBGE, 2020. Disponível em <https://censoagro2017.ibge.gov.br/resultados-censo-agro-2017/cnefe.html>, August 10th, 2022.

INCRA (2022). “Instituto Nacional de Colonização e Reforma Agrária”. <http://www.incra.gov.br/pt/>, August 10th, 2022.

INDE (2022). “Infraestrutura Nacional de Dados Espaciais”. <https://inde.gov.br/>, August.

INPE (2022). “Projeto PRODES: Monitoramento do Desmatamento da Floresta Amazônica Brasileira por Satélite”. <http://www.obt.inpe.br/OBT/assuntos/programas/amazonia/prodes>, August 10th, 2022.

DEEP LEARNING AUTOMATED WORKFLOW FOR CLOUD SEGMENTATION IN REMOTE SENSING IMAGES

Diego Henrique M. Matos¹, Alber H. Sanchez², Tiago G. S. Carneiro³

¹Departamento de Ciência da Computação – Universidade Federal de Minas Gerais (UFMG)
– Belo Horizonte, MG – Brazil

²Instituto Nacional de Pesquisas Espaciais
– São José dos Campos, SP – Brazil

³Departamento de Computação – Universidade Federal de Ouro Preto
– Ouro Preto, MG – Brazil

diegohmm@ufmg.br, alber.ipia@inpe.br, tiago@ufop.edu.br

Abstract. *In this work, we propose an open source and automated workflow for semantic segmentation of remote sensing images. Even though it can be used in other sensors, to evaluate this workflow, a case study has been conducted applying a deep learning algorithm for segmenting clouds in images from WFI sensor, onboard CBERS-4A satellite. Since WFI does not have a tailor-made cloud segmentation algorithm, we customized our workflow based on the U-net neural network to fulfill this gap. Our results are promising according to our tests, although some problems were identified, like false positives over high albedo targets. These problems suggest improvements that could be tackled in the future.*

1. Introduction

A step to quantify and identify deforestation in remote sensing images is identifying clouds. On average, the cloud cover on Earth's surface is between 58% [Rossow and Schiffer 1999] and 66% [Zhang et al. 2004]. When the object of analysis of a remote sensing image is on Earth surface, like forests and cities, the clouds become obstacles between the remote sensor and the object. Thus, detecting and removing clouds are essential steps in remote sensing image analysis [Zhu and Woodcock 2012].

When monitoring large regions, handling data for segmenting clouds is a challenge. This is due to the great number of scenes that need to be processed to form a mosaic covering a region. Each scene is composed by multispectral images, each one containing data from specific spectral bands. Thus, a unique scene implies processing one gigabyte of data. Additionally, each scene has metadata that must be preserved during segmentation process.

In this context, this work develops a deep learning cloud segmentation workflow, which can run on personal computers due to its reduced computational requirements. To accomplish this, scenes are cropped in batches and loaded into memory at the training step, preserving its spatial metadata. This workflow can deal with images from several remote sensors with different characteristics. The only requirement is that input data should be provided in the GeoTIFF file format. Thus, the problem solved in this study is to automatically handle geolocalized images for performing semantic segmentation tasks, allowing the development of novel machine learning-based cloud segmentation algorithms.

The semantic segmentation problem is defined as follows: given an image I , for each pixel X present in the image, determine to which class the pixel belongs. Where i is the image row index, N is the image height, j is the image column index, M is the image width, k is the image band index, and L is the number of the image's channels. Formally, let I be a remote sensing image so that $I = \{X_{ijk} | i = 1, \dots, N, j = 1, \dots, M, k = 1, \dots, L\}$, the semantic segmentation problem is to determine whether the pixel X_{ij} belongs to a determinate class.

In order to evaluate the workflow, this work also conducted a case study. A collection of results obtained from an U-net neural network architecture applied to CBERS-4A WFI remote images is discussed with the goal of evaluating the proposed workflow in cloud segmentation tasks. Images from CBERS-4A Wide Field Camera (WFI) sensor were chosen because they are designed to monitor Amazon rain forest [Souza et al. 2019] and track back deforestation and forest fires. CBERS-4A is capable of making revisits within five days to a certain area [Epiphany 2011]. This feature makes it a good choice for deforestation monitoring tasks.

There are some systems to quantify and identify deforestation using WFI images, like PRODES and DETER [Souza et al. 2019]. However, even disclosing its data, CBERS-4A, a satellite designed by Brazil and China, does not have its own algorithm for cloud semantic segmentation. Currently, INPE uses an algorithm called Cmask [Qiu et al. 2020] to segment clouds [Ferreira et al. 2020]. Nevertheless, Cmask has been originally developed for processing Landsat images and it is not implemented to any remote sensor onboard CBERS-4A. Because of that, the algorithm underperforms. This fact may imply that many images are useless in the processes of forest monitoring.

The rest of this work is organized as follows. Section 2 describes two different branches of the state-of-the-art semantic cloud segmentation and related work. Section 3 presents the methodology, including each step of the proposed workflow, justifying the way they are prepared. Section 4 shows the results of the presented tool applied to a cloud segmentation task. Finally, section 5 presents the conclusion and future work to complement the tool.

2. Related Work

Cloud segmentation could be divided into two large categories: physical - based algorithms and machine learning algorithms. The physical based algorithms use rules based on physical properties, like, for example, the relative angle of sensors to clouds and the angle from the sun, to extract a potential cloud layer and a potential cloud shadow layer [Foga et al. 2017]. This approach also has some disadvantages. First, there are several different remote sensors, with their own spectral bands and resolutions. As a consequence, their physical characteristics are inherent to each sensor and the algorithms can rarely be generalized to more than one sensor. Second, each image goes through the entire algorithm whenever it is segmented, therefore, a physical algorithm, like Fmask [Zhu and Woodcock 2012], can take 0.5 to 6 min to segment a scene on an 8 core Linux server. This makes semantic segmentation a time-consuming task to perform.

Fmask is one of the most used algorithms to segment clouds in Landsat TM images [Zhu and Woodcock 2012]. Currently, it is common to use it for benchmarking other algorithms. It is employed, for example, to evaluate another implementation of Fmask, called CFMask, comparing its performance with "Automated Cloud Cover Assessment" (ACCA) system [Foga et al. 2017].

The Sen2cor algorithm also is a physics-based algorithm, that can classify clouds using just physical data presented in each spectral band [Louis et al. 2016]. This algorithm is created for Sentinel-2 Level-1C products. Cmask, Fmask, and sen2cor algorithms provide evidences on the relationship between physical based algorithms and remote sensor, in a way that the major part of algorithms difficultly adapt to different sensors.

A neural network architecture has been developed allowing the input of remote sensing images of different dimensions, the model is called CloudFCN [Francis et al. 2019]. The mechanism of receive different input images is very useful to allow entering data from more than one remote sensor, thus making the model more generic. Using the Biome dataset [Foga et al. 2017] the model showed an accuracy of 82.81% for RGB scenes and 91.00% with all spectral information versus CFmask [Foga et al. 2017] with 65.69% using all spectral bands available. The lack of spectral bands, using only RGB bands, made it difficult to the model to predict some cases of snow, sand (when it is used remote sensor of high resolution) and other objects that present high albedo.

A new approach for detecting clouds even with hardly distinguishable scenery has been created [Jeppesen et al. 2019]. The model is a neural network built from U-Net and got meaningful results. The "Biome" and "Sparcs" datasets [Foga et al. 2017] were used to evaluate the model divided it into three situations. First, the model is trained with the Biome dataset and evaluated with Sparcs, and vice versa. Second, the Fmask algorithm generates the groundtruth masks. Finally, the model is trained and tested on the same dataset using image-based cross-validation, where the training data are the groundtruth

cloud masks. The model improves the FMask even when using just RGB bands in all the three cited cases of evaluation. This case study has shown that, even with images extracted from a remote sensor with few spectral bands, like CBERS-4A WFI, a deep learning model can produce trustworthy cloud masks.

Currently, the algorithms for semantic segmentation on remote sensing images mainly use deep learning strategies [Yuan et al. 2021]. This occurs because in the last decade, more and more deep learning algorithms have demonstrated good performance in tasks as object classification, performing better than humans in some cases [Lee et al. 2017].

3. Workflows for Cloud Segmentation with Deep Learning Algorithms

The workflow is divided into two main components: model calibration and production model. Workflow components are decoupled from each other. These steps are called in two Jupyter notebook files [Kluyver et al. 2016]. Each notebook implements the main control flow of each component, one for model calibration and another for the production model. The notebook files act as an interface for the user, in which the user may script the workflow in Python programming language [vanRossum 1995], configuring, and connecting steps.

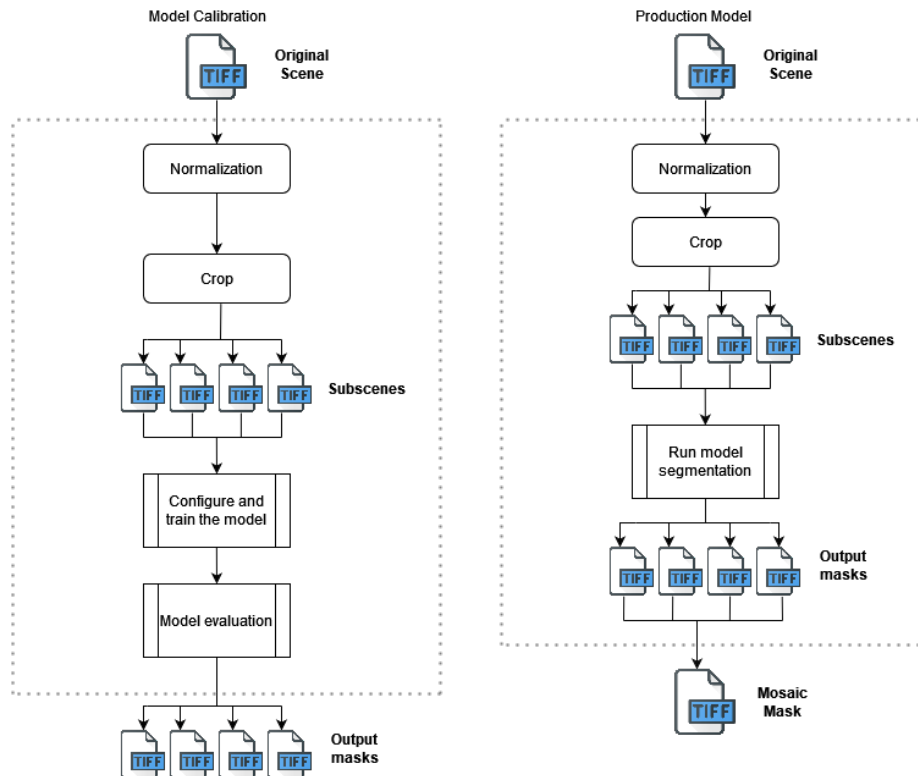


Figure 1. Schema of the proposed workflow. The model calibration (left) is meant to pre-processing the dataset and train the model. The production model (right) is meant to apply new images to be segmented.

The model calibration workflow component consists of several steps: pre-processing remote sensing images (normalization and crop), configuring and training a deep learning model, and compute metrics to evaluate the trained model.

Each step is implemented as a Python function based on the Keras [Ketkar 2017] library that provides features for deep learning model development. The GDAL [Warmerdam 2008] and NumPy [Oliphant 2006] libraries have been used to handle remote sensing images without losing the georeferencing and "no data values" metadata.

3.1. Dataset

Clouds are particularly different entities from objects traditionally segmented in deep learning tasks: the center of a cloud can be dense, with high albedo. However, the border of a cloud can be fuzzy, and can be confused with the background [Dowling and Radke 1990].

In this work, the Cmask algorithm is employed to produce a dataset of segmented images. This strategy allow us to focus on the workflow development. The current dataset is composed by 4 images from WFI remote sensor, that are divided into smaller images, as shown in the figure 2, of 256x256x4 pixels, totaling more than 10.760 images with its respective masks. Since the masks are obtained by applying the Cmask algorithm [Qiu et al. 2020] to each image, they have errors associated and do not correctly represent the ground-truth.

The first step in handling the dataset is to build a directory hierarchy that is divided into the following:

- Original scene folder, to be filled with the original scenes (with all bands);
- normalized scene folder, to be filled with the respective normalized scenes;
- Cropped production folder, to be filled with subscenes from the normalized scenes;
- Cropped production mask folder, to be filled with output masks estimated by the model for each subscene.

In sequence, each scene goes through a Min-Max normalization step, which uses a linear operation to transform the pixel's values [Saranya and Manikandan 2013]. Here, the original pixel's values ranging from 0 to 10.000 are mapped to new values between 0 and 1. The purpose of the normalization step is to transform the image's pixels in a way that each image presents a similar distribution of its pixel values, which helps the neural network optimize its parameters.

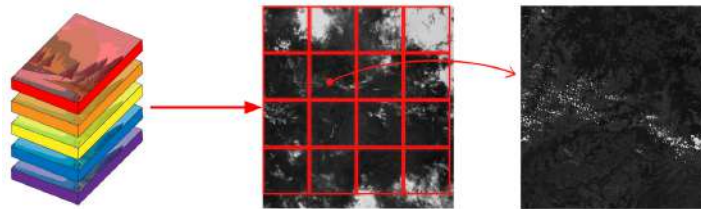


Figure 2. Demonstration of the image cropping step. On the left is the raw scene, in the center is the raw scene with the subscenes in red, on the right is the resulting subscene.

3.2. Deep Learning Model

The chosen neural network has an U-net architecture [Ronneberger et al. 2015], due to its performance and presence in the state-of-the-art. Its architecture is built in two parts. The former is called encoder, and the latter a decoder. The encoder step consists of applying convolutional operations followed by a maxpool downsampling to encode the input image into feature representations at multiple different levels. The decoder step consists of reconstructing the image through upsampling and concatenation at each layer level using the features learned in the encoder step. The encoder-decoder strategy is what makes U-nets good options for the segmentation task, extracting and mapping characteristics from an image.

The tensor representing the subscene can be divided into several dimensions depending on the available memory and the number of bands of the remote sensor. In this study case, each tensor is an image with (256 pixels x 256 pixels x 4 bands). We have chosen a U-net implementation from GitHub with configurable input dimensions [Zak 2021], so that, the resulting workflow can also be configured.

The Keras library contains many metrics available to interpret the confidence of a neural network output, except the F1-Score metric, so it is implemented. It is expected that the user defines a loss function to finish the neural network model. All the loss functions implemented in Keras library are allowed to be used as inputs at this step.

3.3. Training and Test Step

In the learning stage, the model must receive subscenes as input. The set of subscenes are called batches. To load the batches and not use all the machine's memory, a mechanism called "generator" is built to iterate among the images, loading a batch and, after training, reallocating memory to a new batch. In this way, the generator expects the batch size hyperparameter as the input.

After the model consumes the input images and output their respective estimated cloud masks, the ground truth masks will be compared to them. The metrics configured in the previous step are then computed and stored in each pair estimated-ground truth masks.

3.4. Production Model

The first step in handling the dataset is to build a directory hierarchy. The directory folders are divided into the following:

- Production scene folder, to be filled with the original scenes;
- Normalized scene folder, to be filled with the respective normalized scenes;
- Cropped production folder, to be filled with subscenes from the normalized scenes;
- Cropped production mask folder, to be filled with output masks estimated by the model for each subscene;
- Segmented mask folder, to be filled with the entire reconstructed mask.

If there are several scenes in the input dataset, the production model will create an entire directory hierarchy for each scene.

The next step after building the directory hierarchy is to normalize and crop the scenes, a process that is identical to the step in the Model Calibration component.

In the next step, the subscenes are inputted into the pre-trained neural network and the outputs are collected to build a mosaic using the "WARP" function present on GDAL library.

4. Results

The workflow is entirely developed in the Python language and it is executed a 2.60GHz Intel Core i7-9750H machine, with 8 gigabytes of RAM in the Windows 10 operating system. The main employed libraries are Keras, for model implementation; GDAL, to manipulate the georeferenced images; and NumPy, for handle tensors. The following hyperparameters are used:

Table 1. Case study hyperparameter table

Hyperparameter	
Dimension	256
Bands	4
Loss Function	Binary cross-entropy
Optimizer	Adam
Metric	F1-score
Patience	5
Epochs	20

In the case study conducted to evaluate the workflow, it is used a dataset composed by images from CBERS-4A WFI. These images have four spectral bands, Red, Blue, Green and near infra-red. Each scene presented in the training step, must have a ground-truth mask. Cmask algorithm extracted the masks used in this case study [Foga et al. 2017].

In the case study, each subscene is 256x256x4 pixels, being 256 pixels in width and height and 4 layers for red, green, blue, and near infrared spectral bands. The generator mechanism allows the pipeline to work using several images up to 2 gigabytes in size, even with a machine with only 8 gigabytes of RAM.

To choose the loss function parameter the scope of problem must be analyzed. Here, each existent pixel should be classified into two classes: "cloud" or "background." This is a classic case of binary classification, where each pixel in an image either belongs or not to a class. To this scope of problem, the binary cross-entropy, which is based on Bernoulli's formula, is suitable [Ruby and Yendapalli 2020].

The optimizer parameter is a function responsible for minimizing the loss [Géron 2019]. Adam optimizer is one of the most popular and famous gradient descent optimization algorithms [Bock and Weiß 2019]. It is a method that computes adaptive learning rates for each parameter. It stores both the decaying average of the past gradients, similar to momentum and the decaying average of the past squared gradients, similar to RMS-Prop and Adadelta. Thus, it combines the advantages of both methods [Kingma and Ba 2014].

Finally, as the case study has no weight distinction for false positives and false negatives, we chose the F1-score metric to monitor the model learning.

4.1. Performance metrics

The first information generated by the workflow is the history graph represented on the Figure 3.

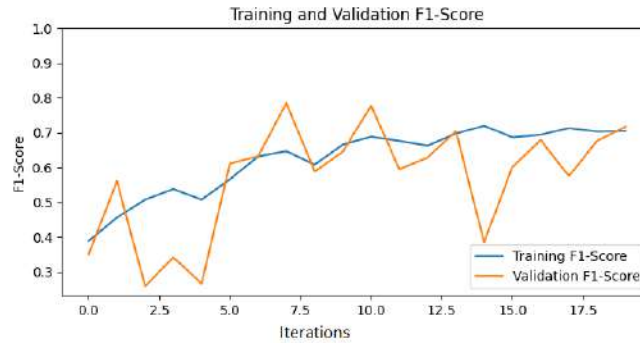


Figure 3. Evolution of the model along the training step.

With this image, it is possible to see metrics in each interaction during the training step, with the F1-score and loss being represented in each epoch. It is possible to observe that, during the training, the model improves at each interaction (Training F1-score) and that it shows a convergence after fifteen interactions. However, not always the validation metrics at the training phase confirm the model performance.

Below, Figure 4 shows some masks from the tested subscenes. The input sub-scenes appear in the left column, the input masks in the middle column, and the predicted mask in the right column.

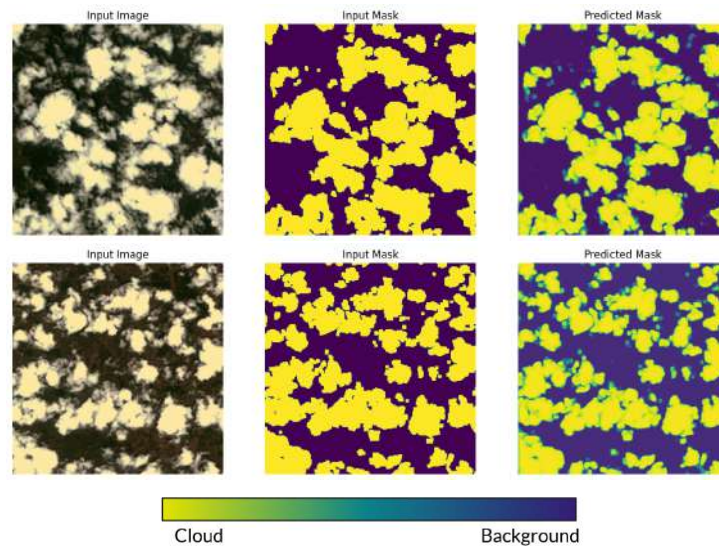


Figure 4. Some input data with the respective predicted masks for the U-net model. Yellow means a high level of probability that the pixel is a cloud. Blue means that there is a low probability that the pixel is a cloud.

The resulting masks present a color gradient between the background color (purple) and the cloud color (yellow), according to the confidence level of the model in

the classification. To make a binary mask, another hyperparameter must be defined, a confidence-level threshold. If the threshold is set to an arbitrarily chosen value such as 0.97, only pixels with 97% confidence will be colored in yellow as clouds. Everything else will be purple. Therefore, it is important to find a threshold value that best avoids wrong coloring. We avoid arbitrarily choosing a threshold value in our case study, preserving the level of confidence analysis.

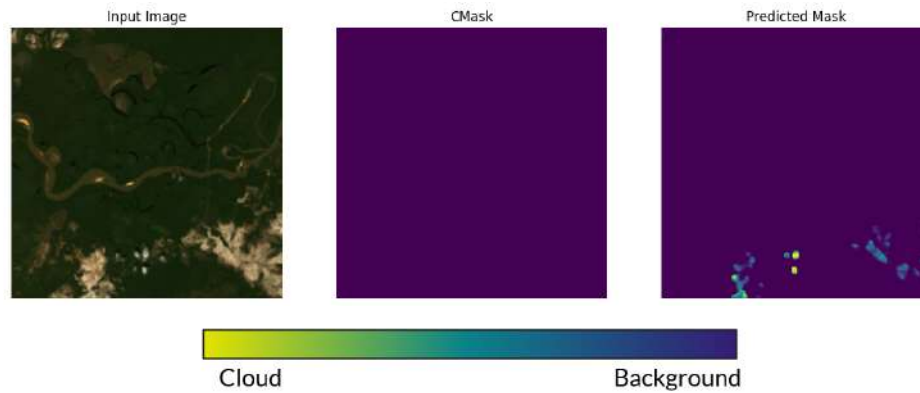


Figure 5. Some input data with the respective predicted mask that present a case of false positive cloudy pixels. Yellow means a high level of probability that the pixel is a cloud. Blue means that there is a low probability that the pixel is a cloud.

As shown in Figure 5, exposed soil has a high albedo compared with vegetation, causing the algorithm to missclassify.

The next step is to assess how the workflow works in a production scenario. After entering the scene in the workflow, the entire scene is segmented and its output mask is reconstructed and saved in the "Merged Files" directory.



Figure 6. Raw Scene.



Figure 7. Mosaic mask.

Figure 7 shows a raw image from CBERS-4A dataset on the left, which the deep learning model has never seen. On the right side, it shows the mosaic of the output-

segmented masks. The mosaic image does not have a defined threshold for pixel confidence levels. In this way, the pixel values are a gradient of values between 0 and 1. By applying a threshold, it is possible to select only pixels where the model has been segmented with high confidence. In Figure 8 no threshold is set, and the Figure 9 a 97% confidence limit is used.



Figure 8. Mosaic mask.



Figure 9. Mosaic mask with threshold.

The two last Figures show the importance of defining a threshold for classifying pixels. This process is called threshold tuning and it requires balancing the recall-precision trade-off [Buckland and Gey 1994]. There are strategies for defining a threshold in segmentation masks, but they are not addressed in this study.

5. Conclusion

This work developed and evaluated a workflow for semantic segmentation of remote sensing images. The proposed workflow handles the dataset automatically, requiring only hyperparameter inputs. The workflow is distributed as a free and open source software, under the LGPL GNU license. Its source code can be found in the repository: <https://gitlab.com/ufopterrallab/projeto-segmentacao-semanticas-cbers-4a-wfi>.

The workflow can be configured for a more powerful machine (increasing sub-scene size and batch size) or running on a personal computer, with less computing power. Even with only 8 gigabytes of RAM, the workflow can train a deep learning model with several images of more than 2 gigabytes each one. The choice of a threshold is excluded from scope of this project, but is left open for future work.

The initial idea of building this workflow was to improve semantic segmentation for the CBERS-4A remote sensor, but in the process of building the dataset and workflow, it became clear that the workflow itself could be a project that would help scientists working with semantic segmentation of remote sensing images, requiring only the hyperparameterization of the model to obtain an initial result.

In the process of building the case study dataset, we identified a scientific and technological gap in cloud removal algorithms for the CBERS-4A WFI remote sensor. Semantic segmentation algorithms that use deep learning appear able to outperform Cmask,

which is currently used by INPE. Additionally, a groundtruth dataset is missing for this remote sensor, limiting the attempts to implement and evaluate other machine learning models to segment could in CBERS-4A WFI images.

References

- Bock, S. and Weiß, M. (2019). A proof of local convergence for the adam optimizer. In *2019 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Buckland, M. and Gey, F. (1994). The relationship between recall and precision. *Journal of the American society for information science*, 45(1):12–19.
- Dowling, D. R. and Radke, L. F. (1990). A summary of the physical properties of cirrus clouds. *Journal of Applied Meteorology and Climatology*, 29(9):970–978.
- Epiphany, J. C. N. (2011). Cbers-3/4: características e potencialidades. In *Proceedings of the Brazilian Remote Sensing Symposium, Curitiba, Brazil*, volume 30, page 90099016.
- Ferreira, K. R., Queiroz, G. R., Vinhas, L., Marujo, R. F., Simoes, R. E., Picoli, M. C., Camara, G., Cartaxo, R., Gomes, V. C., Santos, L. A., et al. (2020). Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24):4033.
- Foga, S., Scaramuzza, P. L., Guo, S., Zhu, Z., Dilley Jr, R. D., Beckmann, T., Schmidt, G. L., Dwyer, J. L., Hughes, M. J., and Laue, B. (2017). Cloud detection algorithm comparison and validation for operational landsat data products. *Remote sensing of environment*, 194:379–390.
- Francis, A., Sidiropoulos, P., and Muller, J.-P. (2019). Cloudfcn: Accurate and robust cloud detection for satellite imagery with deep learning. *Remote Sensing*, 11(19):2312.
- Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. O’Reilly Media.
- Jeppesen, J. H., Jacobsen, R. H., Inceoglu, F., and Toftegaard, T. S. (2019). A cloud detection algorithm for satellite imagery based on deep learning. *Remote sensing of environment*, 229:247–259.
- Ketkar, N. (2017). Introduction to keras. In *Deep learning with Python*, pages 97–111. Springer.
- Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., Ivanov, P., Avila, D., Abdalla, S., and Willing, C. (2016). Jupyter notebooks – a publishing format for reproducible computational workflows. In Loizides, F. and Schmidt, B., editors, *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, pages 87 – 90. IOS Press.
- Lee, J.-G., Jun, S., Cho, Y.-W., Lee, H., Kim, G. B., Seo, J. B., and Kim, N. (2017). Deep learning in medical imaging: general overview. *Korean journal of radiology*, 18(4):570–584.

- Louis, J., Debaecker, V., Pflug, B., Main-Knorn, M., Bieniarz, J., Mueller-Wilm, U., Cadau, E., and Gascon, F. (2016). Sentinel-2 sen2cor: L2a processor for users. In *Proceedings Living Planet Symposium 2016*, pages 1–8. Spacebooks Online.
- Oliphant, T. E. (2006). *A guide to NumPy*, volume 1. Trelgol Publishing USA.
- Qiu, S., Zhu, Z., and Woodcock, C. E. (2020). Cirrus clouds that adversely affect landsat 8 images: what are they and how to detect them? *Remote Sensing of Environment*, 246:111884.
- Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.
- Rossow, W. B. and Schiffer, R. A. (1999). Advances in understanding clouds from isccp. *Bulletin of the American Meteorological Society*, 80(11):2261–2288.
- Ruby, U. and Yendapalli, V. (2020). Binary cross entropy with deep learning technique for image classification. *International Journal of Advanced Trends in Computer Science and Engineering*, 9(10).
- Saranya, C. and Manikandan, G. (2013). A study on normalization techniques for privacy preserving data mining. *International Journal of Engineering and Technology (IJET)*, 5(3):2701–2704.
- Souza, A., Monteiro, A. M. V., Rennó, C. D., Almeida, C. A., Valeriano, D. d. M., Morelli, F., Vinhas, L., Maurano, L. E. P., Adami, M., Escada, M. I. S., et al. (2019). Metodologia utilizada nos projetos prodes e deter. *INPE: São José dos Campos, Brazil*.
- vanRossum, G. (1995). Python reference manual. *Department of Computer Science [CS]*, (R 9525).
- Warmerdam, F. (2008). The geospatial data abstraction library. In *Open source approaches in spatial data handling*, pages 87–104. Springer.
- Yuan, X., Shi, J., and Gu, L. (2021). A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417.
- Zak, K. (2021). keras-unet. <https://github.com/karolzak/keras-unet>.
- Zhang, Y., Rossow, W. B., Lacis, A. A., Oinas, V., and Mishchenko, M. I. (2004). Calculation of radiative fluxes from the surface to top of atmosphere based on isccp and other global data sets: Refinements of the radiative transfer model and the input data. *Journal of Geophysical Research: Atmospheres*, 109(D19).
- Zhu, Z. and Woodcock, C. E. (2012). Object-based cloud and cloud shadow detection in landsat imagery. *Remote sensing of environment*, 118:83–94.

Mapeamento de Corte Seletivo na Amazônia Brasileira Usando Imagens Planet de Alta Resolução Espacial e Inteligência Artificial: estudo de caso na Floresta Nacional do Jamari

Daniel Braga¹, Ricardo Dalagnol^{2,3,4}, Celso B. M. Ribeiro⁵, Liana O. Anderson⁶, Luiz E.O.C. Aragão²

¹Departamento de Geociências – Universidade Federal de Juiz de Fora (UFJF)

Caixa Postal 20010 – 36036-900 – Juiz de Fora – MG – Brazil

²Department of Computer Science – Instituto Nacional de Pesquisas Espaciais (INPE)

São José dos Campos – SP - Brazil

³NASA-Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA 91109, USA.

⁴Center for Tropical Research, Institute of the Environment and Sustainability, University of California, Los Angeles, Los Angeles, CA 90095, USA.

⁵Departamento de Engenharia Sanitária e Ambiental - Universidade Federal de Juiz de Fora (UFJF), Caixa Postal 20010 – 36036-900 – Juiz de Fora – MG – Brazil

⁶ National Centre for Monitoring and Early Warning of Natural Disasters (CEMADEN), São José dos Campos, São Paulo, Brazil

daniel.braga@ich.ufjf.br, ricds@hotmail.com, celso.bandeira@ufjf.edu.br, liana.anderson@gmail.com, luiz.aragao@inpe.br

Abstract. *The Amazon has the largest remnant of tropical forests in the world and is being threatened daily by deforestation and forest degradation. The estimated area of forest degradation is underestimated, which is problematic for both sustainable policy enforcement, environmental oversight and national carbon emission inventories. Therefore, a deep learning model (U-Net) was trained to map forest degradation using Planet imagery (4.77 m spatial resolution) in the Jamari National Forest at the Brazilian Amazon. Preliminary results showed an overall accuracy of 67%. Our approach is promising to monitor forest concessions in Amazonia.*

Resumo. *A Amazônia possui o maior remanescente de florestas tropicais do mundo e está sendo ameaçada diariamente pelo desmatamento e degradação florestal. A área estimada de degradação florestal é subestimada, o que é problemático tanto para a aplicação de políticas sustentáveis, fiscalização ambiental e inventários nacionais de emissões de carbono. Dessa forma, um modelo de deep learning (U-Net) foi treinado para mapear a degradação florestal usando imagens Planet (resolução espacial de 4,77 m) na Floresta Nacional do Jamari, na Amazônia brasileira. Os resultados preliminares mostraram precisão geral de 67%. Nossa abordagem é promissora para monitoramento de concessões florestais na Amazônia.*

1. INTRODUÇÃO

Estima-se que aproximadamente 17% das florestas tropicais desapareceram desde 1990 (VANCUTSEM et al., 2021). Embora a vegetação amazônica ainda abranja oito dos treze países da América do Sul, sua cobertura original sofreu severa redução por conta de processos como o desmatamento e a degradação florestal a partir da década de 1970, principalmente pela inauguração da rodovia Transamazônica (FEARNSSIDE, 2005). Os

processos de degradação afetam a Amazônia em larga escala e já ultrapassam as áreas afetadas por desmatamento (MATRICARDI et al., 2020).

A degradação florestal é caracterizada principalmente por diversos padrões espaciais como: a abertura de estradas na floresta; o corte de árvores; clareiras, tanto das árvores extraídas quanto de outras árvores mortas durante o processo; “cicatrizes” de incêndios; e pátios de extração (PINHEIRO et al., 2016). Dados recentes obtidos do *Tropical Moist Forest* (TMF) do *Joint Research Centre* (JRC) apresentam que a área anual da Amazônia que sofreu com desmatamento e degradação aumentou 24% entre 2019 e 2020 na Amazônia Legal Brasileira, indo de 17.303 km² para 21.379 km² (BEUCHLE et al. 2021). Em adição, as emissões de carbono ainda não são devidamente mensuradas e relatadas nos inventários de emissão de carbono dos países amazônicos (SILVA JUNIOR, et al., 2021).

O objetivo da pesquisa foi avaliar o mapeamento do corte seletivo gerado por um modelo de deep learning para a Floresta Nacional do Jamari, através de imagens de alta resolução espacial Planet. Neste estudo são apresentados os resultados preliminares do mapeamento e a avaliação do desempenho do modelo para mapeamento do corte seletivo relacionado ao manejo florestal.

2. METODOLOGIA

2.1 Área de estudo

A área de estudo foi a Floresta Nacional do Jamari, que pertence à Amazônia Legal e está localizada no norte do estado de Rondônia, entre os municípios de Cujubim, Candeias do Jamari e Itapoã do Oeste (Figura 1). Composta por espécies de árvores de alto valor comercial, sua vegetação ombrófila densa e aberta cobre aproximadamente 2200 km², dos quais cerca de 960 km² (44%) foram inicialmente destinados para a concessão florestal em 2008 (DALAGNOL et al., 2019). É importante reiterar que desde a década de 1980 a área sofre pressão de empreendimentos minerários e da agropecuária, assim como da criação de assentamentos da reforma agrária, o que modificou severamente a paisagem local (BASTOS et al., 2017).

O mapeamento de degradação nesta área seria capaz de identificar processos recorrentes como a abertura de estradas, o corte legal e ilegal, efeitos de borda, queimadas e o próprio desmatamento. As concessões florestais na Amazônia são formas de exploração da floresta na Amazônia de uma forma mais sustentável. Um dos exemplos delas é a Floresta Nacional do Jamari, no estado de Rondônia, a primeira concessão florestal da Amazônia. O monitoramento dessas atividades é importante para garantir que não haja uma exploração além do estabelecido pelo plano de manejo.

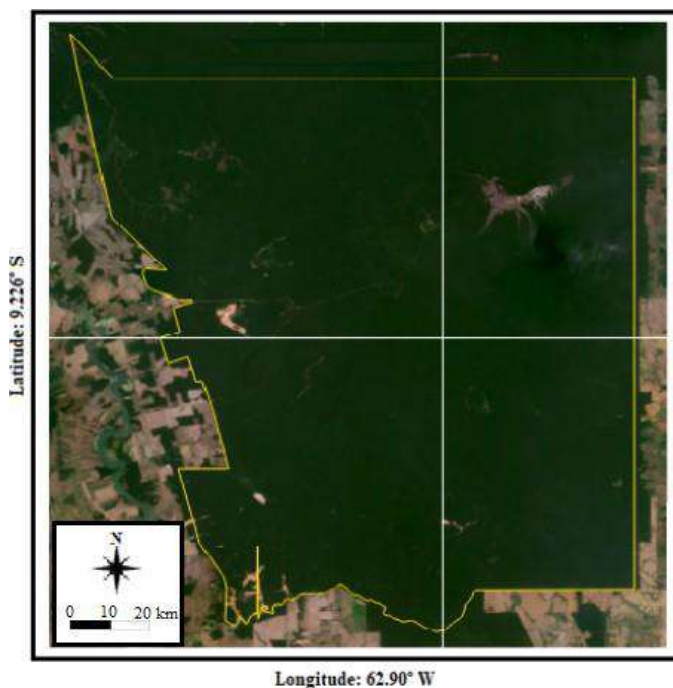


Figura 1 - Limite da área de estudo (Floresta Nacional do Jamari). Imagem no *background* é uma composição colorida cor-verdadeira do satélite Planet produto NICFI.

2.2 Aquisição de imagens Planet e fotointerpretação

O uso de imagens de alta resolução espacial possibilita a fotointerpretação das feições de extração de madeira e queimadas com maior detalhamento. Nesse sentido, a constelação de satélites da missão Planet fornece imagens de alta resolução espacial (~3 – 5 m) e alta revisita (~1 – 3 dias) através do uso de uma constelação (>100) de satélites tipo DOVE (*Digital Orbiting Voice Encoder*). Nesse estudo foram utilizadas as imagens do programa NICFI (*Norway's International Climate and Forest Initiative*) (HERMANSEN et al., 2017), que disponibilizou imagens Planet de alta resolução espacial por meio de um mosaico das florestas tropicais. Foram utilizados mosaicos bi-anuais (6 meses) de 2016 a 2021 para o mapeamento.

Para treinar o modelo, amostras de corte seletivo foram fotointerpretadas na Flona Jamari, bem como em concessões florestais no estado do Mato Grosso. Um total de 289 *quads* ou *tiles* Planet (20x20 km) foram fotointerpretados. Nesse processo, foram utilizadas composições de cor verdadeira (RGB), e o índice de vegetação VARI (Índice Resistente à Atmosfera na Região Visível) (GITELSON et al., 2002). O VARI representa a quantidade de verde na vegetação, calculados a partir de áreas com maior ou menor estresse na vegetação (Equação 1). Ele destaca as áreas com solo exposto ou com dano no dossel da floresta, permitindo mapear as áreas de corte seletivo.

$$VARI = (G - R) / (G + R - B) \quad (1)$$

Equação 1. "VARI" = Índice de Vegetação Resistente à Atmosfera na Região do Visível; "G", "R" e "B" consistem nas bandas do verde, vermelho e azul, respectivamente (Gitelson et al., 2002).

2.3 Modelo Deep Learning

Para realizar a segmentação semântica das áreas de degradação florestal usando imagens Planet, utilizou-se o modelo de redes neurais convolucionais U-Net (RONNEBERGER et al., 2015) aplicado em ambiente R (BUNN, KORPELA, 2014) através do *framework tensorflow* (GOLDSBOROUGH, 2016; RAMPASEK, GOLDENBERG, 2016). O *input* do modelo foram imagens de 256x256 pixels (*patches*) contendo as bandas Planet (R, G, B, NIR) e a máscara de degradação previamente delineada. A partir da amostragem, obteve-se 73.744 *patches* de treinamento. Na arquitetura da U-Net, em cada nível, têm-se duas camadas de convolução com janelas / *kernel* de tamanho 3x3 e um número de atributos que aumenta de acordo com a profundidade da rede até o centro. As camadas convolucionais analisam e aprendem padrões nessas janelas de *pixels* ao mesmo tempo, ao invés de *pixels* individuais. Dessa forma, o modelo é capaz de aprender padrões espaciais relacionados à degradação. O *output* final do modelo é a probabilidade de cada pixel ser caracterizado como extração de madeira. Usando essa probabilidade, obteve-se mapas binários, considerando um limiar maior que 0.5 para áreas afetadas pela exploração. Após treinado, o modelo teve como objetivo realizar a predição de áreas degradadas a partir do treinamento do modelo com padrões espaciais de degradação florestal. O modelo foi utilizado para predizer o corte seletivo anualmente entre 2016 e 2021 para a Flona do Jamari.

2.4 Validação do mapeamento

Para validar o mapeamento das áreas manejadas na Flona do Jamari, foram compilados dados de referência sobre as concessões florestais obtidas no site do Serviço Florestal Brasileiro. Esses dados consistiram em polígonos das áreas manejadas com o ano do manejo (Figura 2). Essas informações foram cruzadas com o mapeamento automático para realizar uma avaliação visual e também quantitativa através de uma matriz de confusão. Calculou-se a acurácia global entre o mapeamento e a referência. Foi calculada a área total mapeada como corte seletivo dentro da Flona Jamari, para áreas dentro e fora das concessões florestais.

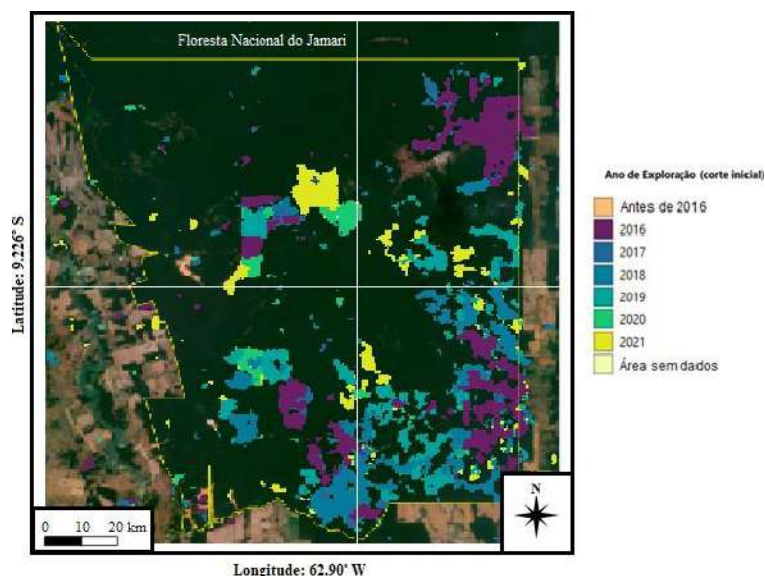


Figura 2 - Unidades de produção (UPA) de concessões florestais na Flona Jamari. Fonte de dados: Serviço Florestal Brasileiro, 2022.

3. RESULTADOS E DISCUSSÃO

3.1 Mapeamento

O modelo gerou um mapa de corte seletivo para a Floresta Nacional do Jamari durante o período de 2016 a 2021 (Figura 3). Nesse mapa, foi possível identificar a localização da extração de madeira por toda extensão da Flona, de acordo com a data em que ocorreu o corte segundo a disponibilidade das imagens Planet. Cada cor representa o ano em que o corte foi detectado pelo modelo. As 9 UPAs com data de corte na referência oficial entre 2016 e 2019 foram corretamente identificadas pelo modelo na mesma data do corte observado em campo, ou com alguma diferença mensal devido ao fato de as imagens Planet geradas antes de Setembro de 2020 serem mosaicos semestrais. As UPAs com ano de corte anterior à 2016 o modelo não foi capaz de identificar, uma vez que as imagens Planet ainda não estavam disponíveis.

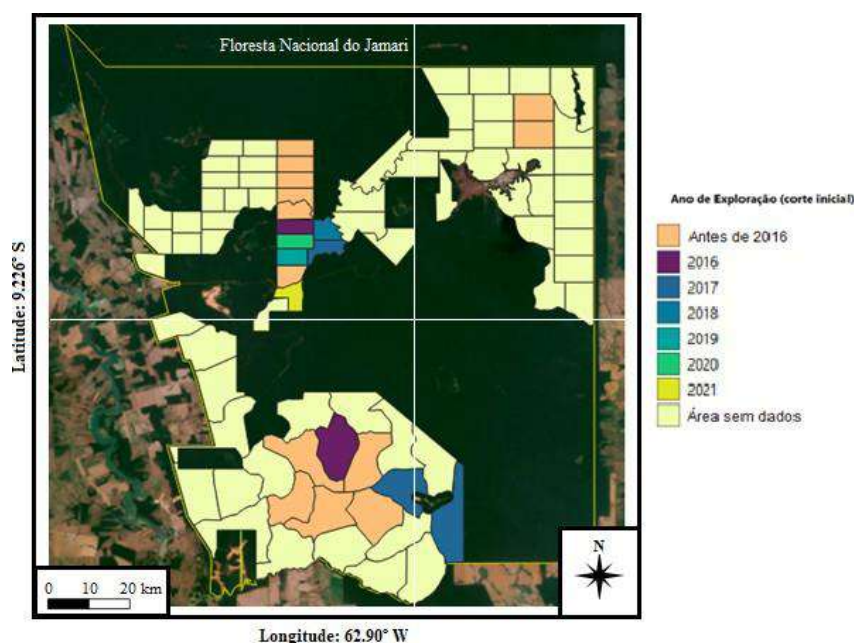


Figura 3 – Mapa de predição de corte seletivo na Flona do Jamari. Cada cor corresponde a um ano de corte.

Como esperado, o modelo não detectou as UPAs com data de corte na referência oficial anterior a 2016, uma vez que o imageamento se inicia a partir de dezembro de 2015.

O modelo detectou a presença significativa de padrões de corte seletivo em 27 UPAs que não possuem nenhuma referência do SFB, que estão na categoria “Área não explorada / Sem dados” (Figura 4). A UPA da UMF I apresenta padrão de corte semelhante ao das UPAs com referência, as UPAs da UMF II apresentam padrão divergente do que foi observado no corte seletivo e as UPAs da UMF III variam entre as duas opções. O modelo também detectou áreas de corte seletivo dentro do perímetro de UPAs que não tinham informação de corte. Essas áreas apresentaram padrões de corte seletivos semelhantes ao de UPAs com referência. Esses padrões consistiram na ocorrência de trilhas regulares, pátios para estocagem de madeira espaçados, e clareiras no dossel da floresta.

O modelo detectou áreas de corte seletivo fora do perímetro de UPAs sem referência oficial do SFB, com padrões de corte diferente do observado no corte seletivo dentro das UPAs. Esses padrões consistiram na abertura de diversas trilhas desordenadas dentro da floresta na parte sudoeste, e sem ocorrência regular de pátios para estocagem da madeira. Essas áreas provavelmente consistem em cortes ilegais.

3.2 Análise quantitativa do mapeamento

Para uma análise estatística do mapeamento, foi gerada uma matriz de confusão (Tabela 1) a partir do cruzamento entre a predição do modelo U-Net e a referência do SFB. Os números representam a quantidade de pixels encontrados entre as classes de referência e predição. A partir da matriz de confusão, foi calculada uma exatidão global de 67%. Isso quer dizer que 67% da área foi corretamente mapeada de acordo com a referência do SFB.

Como apontado na análise visual, observamos que existem áreas que apresentaram padrões claros de corte seletivo, porém não tinham referência do ano do corte. Esse é o caso dos valores observados na coluna “0” (não detecção) onde são encontrados muitos pixels com predição entre 2016 até 2021. Uma parte dessas áreas ocorre por conta de possíveis inconsistências na referência do SFB que devem ser exploradas. Essa divergência reduz a exatidão global do modelo.

		Referência						
		0	2016	2017	2018	2019	2020	2021
Predição	0	6354	47	198	16	0	3	24
	2016	688	218	110	2	4	1	6
	2017	139	38	43	8	0	10	0
	2018	583	0	119	17	0	0	0
	2019	500	0	46	2	66	42	0
	2020	225	0	0	0	0	14	49
	2021	427	0	0	0	0	0	1

Tabela 1 – Matriz de Confusão entre a predição da extração de madeira e a referência oficial das datas de corte SFB. Nas linhas e colunas constam o ano da detecção ou valor=0, que corresponde à ausência do corte seletivo. As unidades correspondem a pixels Planet (4.77m).

O modelo identificou a presença de corte seletivo dentro do perímetro da UMFs e também fora destas, o que evidenciou perturbações relacionadas à degradação que não constam no banco de dados das concessões florestais do SFB. Foram detectados 513 km² (23,31%) de floresta com detecção de corte seletivo na Flona Jamari entre 2016 e 2021. Desse total de detecções, 58% foram identificados dentro das zonas de manejo, num total de 298 km², ou seja, 13,54% da área total da Flona.

4. CONSIDERAÇÕES FINAIS

O modelo aqui avaliado se mostrou promissor para identificação do corte seletivo de uma concessão florestal na Amazônia com uma exatidão global de 67%. Ele detecta padrões de corte seletivo, com abertura de estradas, pátios de estocagem de madeira e clareiras geradas pelo corte das árvores. Os resultados apresentam divergências nas métricas de avaliação estatísticas, que pode ser devido a inconsistências nos dados usados como referência.

Há dois tipos de áreas em que o modelo detectou a presença de degradação possivelmente ilegal: áreas de corte ilegal dentro das UPAs; e áreas de corte ilegal fora das UPAs. No primeiro caso, a extração ocorre aparentemente de forma desordenada, sugerindo

possível corte ilegal dentro das UPAs, uma vez que o padrão está diferente das UPAs que possuem referência de dados de corte. No segundo caso, o modelo detectou padrões de degradação e corte ilegal fora das UPAs, o que, assim como no primeiro caso, pode representar um sério problema na fiscalização e monitoramento das florestas públicas e nas concessões florestais. Sendo assim, por mais que o modelo detecte corretamente os padrões de degradação relacionados ao desmatamento, ainda é preciso validar algumas informações junto ao SFB, principalmente para a Floresta Nacional do Jamari.

5. REFERÊNCIAS

- BASTOS, A.; MANIESI, V.; GOMES, F. B.. Evolução do desmatamento e implicações ao uso e ocupação dos solos no entorno de unidades de conservação da Amazônia – o caso da floresta nacional do Jamari/RO. Os Desafios da Geografia Física na Fronteira do Conhecimento, p. 6612–24, 2017.
- BEUCHLE, R.; ACHARD, F.; BOURGOIN, C.; VANCUTSEM, C.; EVA, H. D.; FOLLADOR, M. Deforestation and Forest Degradation in the Amazon – Status and Trends up to Year 2020, EUR30727EN, Publications Office European Union, Luxembourg, 2021.
- BUNN, A.; KORPELA, M. Crossdating in dplR. 2014.
- DALAGNOL, R.; PHILLIPS, O. L.; GLOOR, E.; GALVÃO, L. S.; WAGNER, F. H.; LOCKS, C. J.; ARAGÃO, L. E. O. C.. Quantifying Canopy Tree Loss and Gap Recovery in Tropical Forests under Low-Intensity Logging Using VHR Satellite Imagery and Airborne LiDAR. Remote Sensing, v. 11, n. 7, p. 817, 2019.
- FEARNSIDE, P. M.. Deforestation in Brazilian Amazonia: history, rates, and consequences. Conservation Biology, v. 19, n. 3, p. 680-688, 2005.
- GITELSON, A. A.; ZUR, Y.; CHIVKUNOVA, O. B.; MERZLYAK, M. N.. Assessing Carotenoid Content in Plant Leaves with Reflectance Spectroscopy. Photochemistry And Photobiology, v. 75, n. 3, p. 272, 2002.
- GOLDSBOROUGH, P. A Tour of Tensorflow. 2016.
- HERMANSSEN, E.; MCNEILL, D.; KASA, S.; RAJÃO, R.. Co-Operation or Co-Optation? NGOs' Roles in Norway's International Climate and Forest Initiative. Forests, v. 8, n. 3, p. 64, 2017.
- MATRICARDI, E. T.; SKOLE, D. L.; COSTA, O. B.; PEDLOWSKI, M. A.; SAMEK, J. H.; MIGUEL, E. P.. Long-term forest degradation surpasses deforestation in the Brazilian Amazon. Science, v. 369, n. 6509, p. 1378-1382, 2020.
- PINHEIRO, T. F.; ESCADA, M. I. S.; VALERIANO, D. M.; HOSTERT, P.; GOLLNOW, F.; MÜLLER, H.. Forest Degradation Associated with Logging Frontier Expansion in the Amazon: the BR-163 region in southwestern Pará, Brazil. Earth Interactions, v. 20, n. 17, p. 1-26, 2016.
- RAMPASEK, L.; GOLDENBERG, A. TensorFlow: biology's gateway to deep learning?. Cell Systems, v. 2, n. 1, p. 12-14, 2016.
- RONNEBERGER, O.; FISCHER, P.; BROX, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Em: NAVAB, N. et al. (Eds.). M I Computing and Comp-Assist Int – MICCAI. Springer International Publishing, v. 9351 p. 234–241, 2015.
- SILVA JUNIOR, C. H. L. et al. Amazonian Forest Degradation Must be Incorporated into the COP26 Agenda. Nature Geoscience, v. 14, n. 9, p. 634–635, 2021.
- VANCUTSEM, C.; ACHARD, F.; PEKEL, J.-F.; VIEILLEDENT, G.; CARBONI, S.; SIMONETTI, D.; GALLEGGO, J.; ARAGÃO, L. E. O. C.; NASI, R. Long-term (1990–2019) Monitoring of Forest Cover Changes in the Humid Tropics. Science Advances, 7 (10), 2021.

Um algoritmo de máxima concordância para harmonizar legendas de mapas de uso e cobertura da terra

Sabrina G. Marques¹, Pedro R. Andrade¹, Aline C. Soterroni²

¹Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil

²Department of Biology, University of Oxford, Oxford, UK

sabrina.marques@inpe.br, pedro.andrade@inpe.br,
aline.soterroni@biology.ox.ac.uk

Abstract. *The growing production of land use and land cover maps has made it easier to monitor land changes. However, the integration and harmonization of maps can be a challenging task, since this diversity of initiatives also implies in a diversity of legends and class definitions. This article proposes an algorithm, based on the concept of maximum agreement, to help mapping legends from land use and land cover maps. The proposed algorithm was used to compare maps produced by TerraClass and MapBiomass in a small region of Tocantins. The results indicate that the automated production of a maximum agreement legend between two maps can be a feasible initial step in the legend harmonization process, being performed even before any semantic mapping between classes.*

Resumo. *A crescente produção de mapas de uso e cobertura da terra facilitou o monitoramento das mudanças de uso e cobertura do solo. No entanto, harmonizar e integrar esses mapas pode ser uma tarefa desafiadora, uma vez que essa diversidade de iniciativas também implica em uma diversidade de legendas e classes. Esse artigo propõe um algoritmo, baseado no conceito de concordância máxima, para ajudar no mapeamento de legendas de diferentes mapas de uso e cobertura da terra. O algoritmo proposto foi usado para comparar mapas produzidos pelo TerraClass e MapBiomass em uma pequena região do Baixo Tocantins. Os resultados indicam que a produção automatizada de uma legenda de concordância máxima entre dois mapas pode ser um passo inicial viável no processo de harmonização de legendas, sendo executado antes mesmo de qualquer mapeamento semântico entre as classes.*

1. Introdução

As atividades humanas provocam impactos na superfície da terra, sendo uma das principais forças nas alterações do sistema terrestre e no clima do planeta. Em escala global, o setor de agricultura, florestas e outros usos da terra é responsável por quase um quarto das emissões de gases de efeito estufa [Shukla et al. 2019]. Nesse contexto, os mapas de Uso e Cobertura da Terra (UTC) podem ser usados para acompanhar as dinâmicas do solo, bem como para estudar as interações sociais que afetam esses processos.

Com os avanços tecnológicos e o aumento significativo na disponibilidade de dados geoespaciais, temos como resultado diversas iniciativas produzindo mapas de UCT regionais e globais. No entanto, harmonizar e integrar mapas de UCT é uma tarefa desafiadora, uma vez que essa diversidade de iniciativas e dados disponíveis também implica em uma diversidade de metodologias, produtos e legendas que são desenvolvidos de

acordo com os objetivos dos estudos e as aplicações desejadas. Com isso, nos deparamos com uma variedade de sistemas de classificação com diferentes semânticas para as classes propostas. Por melhor que sejam esses mapeamentos, essas diferenças podem indicar inconsistências ao comparar mapas de UCT de iniciativas diferentes, ou até mesmo para versões distintas de uma mesma iniciativa [Verburg et al. 2011, Köhl et al. 2000].

O processo usual para comparar mapas de UCT consiste em elaborar uma legenda harmonizada a partir das descrições semânticas de cada uma das classes apresentadas pelos dois mapas [Neves et al. 2017, Reis et al. 2018, Capanema et al. 2019]. Dessa forma, algumas classes são ignoradas por falta de concordância ou informação. Outras são agrupadas em uma classe maior e mais genérica a fim de diminuir as dissimilaridades entre os conjuntos. Sistemas de classificação também costumam ser usados para traduzir essas diferenças semânticas em uma nova legenda igual para ambos os mapas, possibilitando compará-los através de diferentes métricas. O *Land Cover Classification System* (LCCS) é um dos mais relevantes sistemas de classificação e foi proposto pela *Food and Agriculture Organization* (FAO) com intuito de fornecer uma estrutura consistente para a classificação e mapeamento hierárquico da cobertura da terra. Ele foi desenvolvido para padronizar as terminologias dos atributos usados a partir de um conjunto de critérios independentes [Gregorio 2016].

Abordagens usando dados extraídos dos próprios mapas, considerando a distribuição espacial dessas classes, não são usuais. Este trabalho assume como hipótese que realizar uma análise prévia sobre esses dados pode auxiliar no processo de harmonização de legendas, uma vez que pode fornecer informações tanto sobre classes cujo mapeamento é trivial quanto sobre dubiedades e confusões no mapeamento. Nesse sentido, o objetivo deste trabalho é propor um algoritmo de harmonização automatizada de legendas que produza a maior concordância possível entre dois mapas.

2. Metodologia

Esta seção apresenta o algoritmo de harmonização de legendas de dois mapas de UTC. A ideia do algoritmo é prover informações que auxiliem o usuário a realizar o mapeamento entre as classes de dois mapas de uso da terra. Ele computa uma concordância máxima entre os dois mapas, o que significa que qualquer mapeamento de classes diferente do escolhido pelo algoritmo irá produzir uma concordância menor ou igual à concordância resultante do algoritmo.

O algoritmo recebe como entrada dois mapas de uso da terra M_1 e M_2 e retorna ao usuário essa proposta de legenda de acordo com as concordâncias identificadas. M_1 possui as classes $x_1, x_2, \dots, x_m$, e M_2 possui as classes $y_1, y_2, \dots, y_n$. O algoritmo assume que as representações espaciais dos dois mapas são compatíveis. É necessário que ambos mapas estejam na mesma projeção, em formato matricial, possuindo o mesmo número de linhas e colunas e tenham a mesma resolução espacial. Caso não estejam, é necessária uma etapa de pré-processamento desses mapas que não é considerada pelo algoritmo.

O algoritmo possui 3 passos. O primeiro passo consiste em produzir uma tabulação cruzada entre os mapas, dada pela matriz A onde o valor da posição a_{ij} da matriz representa a contagem de *pixels* de concordância entre a classe x_i de M_1 e a classe y_j de M_2 . Ela é da forma:

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}, A \in N^{m \times n}, \quad (1)$$

No segundo passo são determinadas as concordâncias por linha e por coluna a partir de A , formando dois conjuntos, C_L e C_C , a partir das funções

$$c_l(x_i) = \{y_k : n_l(a_{ik}) = \max_{p=1, \dots, n} \{n_l(a_{ip})\}\}, \text{ onde} \quad (2)$$

$$n_l(a_{ij}) = \frac{a_{ij}}{\sum_{p=1}^n a_{ip}}, \text{ e} \quad (3)$$

$$c_c(y_i) = \{x_k : n_c(a_{ki}) = \max_{p=1, \dots, m} \{n_c(a_{pi})\}\}, \text{ onde} \quad (4)$$

$$n_c(a_{ij}) = \frac{a_{ij}}{\sum_{p=1}^m a_{pj}}, \text{ respectivamente.} \quad (5)$$

O conjunto C_L contém m duplas, $(x_i, c_l(x_i))$, $i \in \{1, \dots, m\}$ formadas pelas classes de M_1 as respectivas classe de M_2 que apresentaram maiores concordâncias de acordo com a matriz A . De forma análoga, o conjunto C_C é formado por n duplas, $(c_c(y_i), y_i)$, $i \in \{1, \dots, n\}$, da concordância das classes de M_2 em relação às classes de M_1 , de acordo com A .

Após formados os dois conjuntos, o último passo do algoritmo consiste em produzir a legenda harmonizada que será dada por

$$C_L \cup C_C, \quad (6)$$

contendo todas as concordâncias obtidas pelas funções, c_l e c_c . Esse conjunto poderá ter como tamanho máximo a soma das quantidades de classes dos dois mapas e como tamanho mínimo a quantidade de classes do mapa com mais classes.

Essa união entre os conjuntos C_L e C_C engloba três possibilidades de mapeamentos entre as classes dos mapas. A primeira é a mais simples, ocorrendo quando existe um mapeamento por linha da classe x_i para a classe y_j e também existe um mapeamento por colunas entre y_j e x_i no segundo mapeamento, assim a dupla (x_i, y_j) é considerada na harmonização. O segundo caso ocorre quando o mapeamento de uma classe existe somente em uma das harmonizações. Por exemplo, quando existe um mapeamento por linha da classe x_i para a classe y_j , mas a classe x_i não é mapeada na harmonização por colunas. Nesse caso a dupla (x_i, y_j) é considerada na harmonização. No último caso, existe um mapeamento por linha da classe x_i para a classe y_j mas no mapeamento por coluna a classe y_j é mapeada com a classe x_k , $k \neq i$, indicando uma inconsistência. De qualquer forma, as duplas (x_i, y_j) e (x_k, y_j) serão consideradas na legenda harmonizada.

3. Estudo de Caso

Como estudo de caso, o algoritmo proposto foi aplicado em uma pequena região do baixo Tocantins, que compreende os municípios paraenses de Mocajuba, Baião e Cametá. Esse

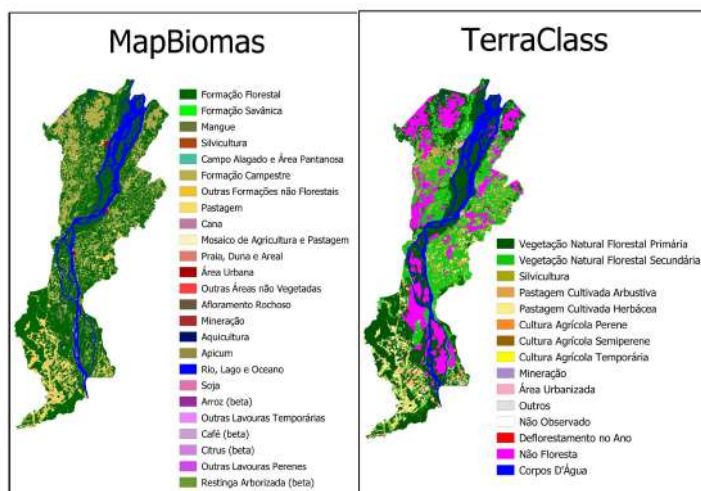


Figura 1. Mapas de Uso e Cobertura da Terra da região de interesse para 2014.

local foi escolhido com o intuito de replicar o trabalho feito por [Capanema et al. 2019], que realizou uma comparação entre o TerraClass Amazônia e o MapBiomas Coleção 2, ambos para o ano de 2014. Como os dados da Coleção 2 não se encontram mais disponíveis, neste estudo foram usados os dados do MapBiomas Coleção 6. A Figura 1 mostra os mapas de UCT de ambas as iniciativas para a área de estudo.

A Figura 2 mostra as equivalências por linha e coluna obtidas pelo algoritmo. As linhas azuis mostram as classes que tiveram a mesma equivalência nas harmonizações, ou seja, na harmonização por linha e por coluna foi formada a mesma dupla de classes. Como exemplo, temos o caso das classes de ‘Formação Florestal’ do MapBiomas e ‘Vegetação Natural Florestal Primária’ do TerraClass que foram mapeadas tanto na harmonização por linhas, quanto por colunas. As linhas laranjas representam classes que tiveram diferentes equivalências. Na harmonização por linha temos que a classe ‘Outras Lavouras Perenes’

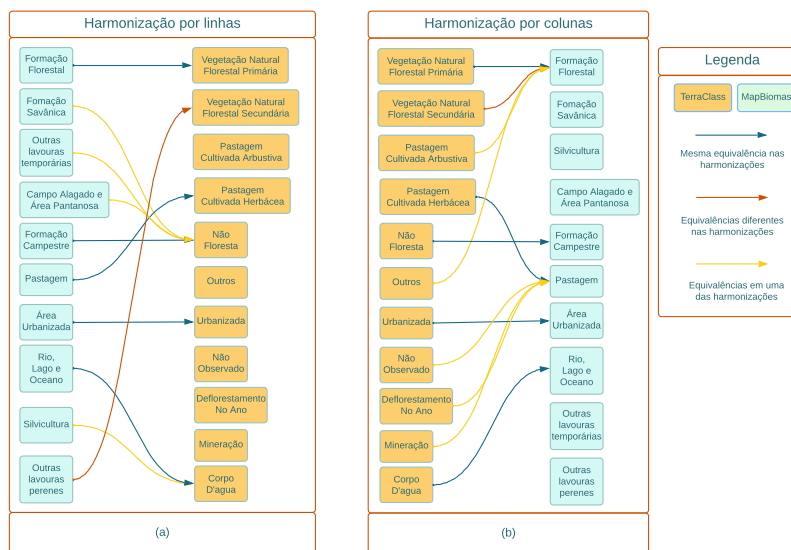


Figura 2. Equivalências por linha e coluna.

foi mapeada como ‘Vegetação Natural Florestal Secundária’, mas essa classe foi mapeada como ‘Formação Florestal’ na harmonização por colunas. Em amarelo estão as classes que foram identificadas em apenas uma das equivalências, como o caso da classe ‘Pastagem Cultivada Arbustiva’ do TerraClass que foi mapeada como ‘Formação Florestal’ na harmonização por colunas, mas na harmonização por linhas não houve mapeamento.

É importante ressaltar que nesse estudo de caso ocorreram as três possibilidades de mapeamento discutidas na metodologia. O segundo e o terceiro caso são aqueles em que o pesquisador deverá se atentar ao comparar os mapas ao fazer a harmonização das legendas, pois é onde as maiores divergências acontecem e, eventualmente, pode ser necessário desconsiderar algumas dessas classes na análise a ser realizada.

Por fim, após produzidos esses conjuntos de harmonizações, eles são agrupados formando a legenda harmonizada. A Figura 3 apresenta a proposta de legenda harmonizada, baseada nas equivalências por linha e por coluna obtidas. É possível observar que a legenda ótima produzida pode apresentar algumas inconsistências, como o mapeamento entre a classe de ‘Silvicultura’ no MapBiomias e ‘Corpo D’água’ no TerraClass (note que ‘Corpo D’água’ no TerraClass foi mapeado como ‘Rio, Lago e Oceano’ no MapBiomias), o que fornece uma evidência das semânticas adotadas pelos diferentes mapeamentos.

A concordância entre os mapas usando o mapeamento produzido pelo algoritmo foi de 78%, enquanto que [Capanema et al. 2019] obtiveram uma concordância de 52% para a mesma região com uma versão anterior do MapBiomias. Esse aumento pode estar associada à melhor qualidade das versões mais recentes do MapBiomias. Em [Neves et al. 2017], a concordância entre essas iniciativas para o estudo do Pará foi de 84%, enquanto que em [Neves et al. 2020] a concordância foi de 87% e 92% para a região do bioma Amazônia, entre diferentes versões dessas iniciativas. Ambos os estudos apresentaram um concordância maior do que aquela apresentada pelo algoritmo, no entanto tratam-se de análises em regiões maiores, de forma que as diferenças locais podem ter impacto significativo na concordância.

Esta análise preliminar possibilita uma visão mais ampla das classes de uma le-



Figura 3. Legenda harmonizada obtida pelo algoritmo.

genda que não possuem equivalência direta com outra, como é o caso da classe ‘Mosaico de Agricultura e Pastagem’ do MapBiomias, que não foi identificada na área de estudo considerada. Caso fosse, seria possível analisar a qual classe ela está majoritariamente mapeada, possivelmente remapeando ao invés de desconsiderar em um estudo.

4. Conclusão

O algoritmo proposto produz uma harmonização automática entre legendas de mapas mapas de UCT que possui a maior concordância entre os dois mapas. Adicionalmente, ela poderá indicar eventuais inconsistências entre classes e fornecer uma análise das classes que não possuem uma equivalência direta com outra legenda. Esse algoritmo pode então ser um primeiro passo automatizado para o processo de mapeamento de classes. Ao invés de iniciar o processo de mapeamento das classes apenas a partir da semântica das legendas, o usuário poderá elaborar uma legenda usando como suporte esse mapeamento baseado na concordância das classes. Vale ressaltar que qualquer legenda diferente desta terá uma concordância menor ou igual à obtida pelo algoritmo proposto neste trabalho.

Os resultados obtidos no estudo de caso se mostraram promissores e indicam que a produção automatizada de uma legenda ótima pode trazer benefícios para a área de UCT. O código do experimento desse artigo está disponível em um *notebook*¹. Como trabalho futuro será desenvolvido um pacote *Python* com o algoritmo proposto neste trabalho e um *plugin* para o QGIS com intuito de facilitar o seu uso.

Referências

- Capanema, V. et al. (2019). Comparação entre os produtos temáticos de uso e cobertura da terra do terraclass amazônia e mapbiomas: Teste de aderência entre classes. *Anais do XIX Simposio Brasileiro de Sensoramento Remoto*, pages 724–727.
- Gregorio, A. D. (2016). *Land Cover Classification System: Classification concepts*. Food and Agriculture Organization of the United Nations (FAO), Rome.
- Köhl, M. et al. (2000). Harmonisation and standardisation in multi-national environmental statistics – mission impossible? *Environmental Monitoring and Assessment*, 63:361–380.
- Neves, A. K. et al. (2017). Terraclass x mapbiomas: Comparative assessment of legend and mapping agreement analysis. *Proceedings of the Brazilian Symposium on Geoinformatics*, 2017-Decem:295–300.
- Neves, A. K. et al. (2020). Assessment of terraclass and mapbiomas data on legend and map agreement for the Brazilian amazon biome. *Acta Amazonica*, 50:170–182.
- Reis, M. S. et al. (2018). Towards a reproducible LULC hierarchical class legend for use in the southwest of Pará state, Brazil : Data-driven hierarchies. *Land*, 7.
- Shukla, P. R. et al. (2019). Climate change and land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems.
- Verburg, P. H. et al. (2011). Challenges in using land use and land cover data for global change studies. *Global Change Biology*, 17(2):974–989.

¹<https://www.kaggle.com/code/sabrinamarques/algoritmo-de-harmoniza-o-estudo-de-caso>

Acessibilidade a Escolas na Região do Grande ABC: Desigualdades Socioespaciais

Roberta P. de Magalhães¹, Flávia da F. Feitosa¹, Diego B. Tomasiello^{1,2}

¹Centro de Engenharia, Modelagem e Ciências Sociais Aplicadas (CECS) e Laboratório de Estudos e Projetos Urbanos e Regionais (LEPUR) – Universidade Federal do ABC (UFABC)

²Centro de Estudos da Metrópole (CEM) e Laboratório de Geoprocessamento – Universidade de São Paulo (USP)

robertaperez.magalhaes@gmail.com, flavia.feitosa@ufabc.edu.br,
diegobt86@gmail.com

Abstract. *This study explores the spatial inequalities of access to schools through walking. Two accessibility metrics – Cumulative and Balanced Floating Catchment Area (BFCA) – were applied in the Greater ABC Region, São Paulo, to understand their potential and limitations. It was observed that although the cumulative metric does not consider the competition between opportunities, while the BFCA considers it, the spatial results of the two metrics were complementary in the studied region. Furthermore, the two metrics showed that the correlation between accessibility and income is negative for lower-income groups and positive for higher-income groups. Thus, it was found that spatial analysis from the use of such measures can be an important tool to support the formulation of public policies territorially oriented.*

Resumo. *Este estudo explora as desigualdades espaciais de acesso às escolas através da caminhada. Duas métricas de acessibilidade – Cumulativa e Balanced Floating Catchment Area (BFCA) - foram aplicadas na Região do Grande ABC, São Paulo, a fim de compreender suas potencialidades e limitações. Observou-se que apesar de a métrica cumulativa não levar em conta a competição entre oportunidades, ao passo que a BFCA a considera, os resultados espaciais dos dois índices foram complementares na região estudada. Ademais, as duas métricas apontaram que a correlação entre acessibilidade e renda é negativa para grupos de menor renda e positiva para grupos de maior renda. Assim, constatou-se que a análise espacial a partir do uso de tais medidas pode ser uma importante ferramenta para subsidiar a formulação de políticas públicas territorialmente orientadas.*

1. Introdução

A partir do século XX, o processo brasileiro de urbanização experimentou uma rápida aceleração. Esse crescimento urbano, entretanto, não ocorreu de forma homogênea territorialmente, o que acabou por gerar desigualdades espaciais em diversos âmbitos: sociais, econômicos, urbanos, ambientais etc. Nessa conjuntura, a partir da década de 1980, as periferias passaram a crescer em ritmo maior que os próprios núcleos urbanos metropolitanos, caracterizando-se como o destino da população pobre que não conseguia estabelecer-se nas áreas centrais da cidade devido aos altos custos. Configurou-se, assim, uma característica intrínseca às metrópoles brasileiras: a segregação socioespacial (Maricato, 2003).

Dessa forma, a acessibilidade representa uma variável imprescindível nos estudos das cidades. Definida como a facilidade com que um indivíduo pode acessar atividades dispostas ao longo do espaço através de um ou alguns modais de transporte, a acessibilidade é capaz de relacionar o modelo de uso e ocupação da terra e o sistema de transportes da área de estudo. A acessibilidade pode também ser compreendida como um indicador social, já que suas métricas são capazes de apontar as desigualdades de acesso a oportunidades específicas entre grupos populacionais definidos a partir do objetivo da análise. Assim, os índices de acessibilidade são potenciais elementos de subsídio para formulação de políticas públicas (Hansen, 1959; Geurs e Van Wee, 2004).

Nesse contexto, este trabalho buscou utilizar duas métricas distintas de acessibilidade a fim de explorar as desigualdades de acesso através de caminhadas com duração de no máximo 15 minutos até escolas de nível fundamental e médio na Região do Grande ABC (RGABC), localizada na Região Metropolitana de São Paulo (RMSP), e, por fim, analisar suas correlações com a distribuição espacial de grupos de renda. A primeira métrica computada é a que recebe o nome de “Cumulativa”, já que é resultado da acumulação quantitativa de oportunidades que a população residente nas unidades espaciais analisadas pode acessar dentro de certo limite de tempo. A segunda, por sua vez, é chamada “Balanced Floating Catchment Areas” (BFCA), responsável por captar a demanda para algum tipo de oportunidade e relacioná-la com o nível de serviço (oferta) que cada uma das potenciais oportunidades alcançadas dentro do limite de tempo estabelecido são capazes de oferecer. Em síntese, foi possível atestar que apesar de a métrica cumulativa de acessibilidade possuir uma interpretação mais simples, ela não considera a competição existente entre as oportunidades, que nem sempre comportam a demanda existente. Em contraposição, a métrica BFCA considera tal competição. Ademais, observou-se, a partir do coeficiente de correlação de Spearman, que grupos sociais de menor renda tendem a residir em áreas onde a acessibilidade a escolas – diante das duas métricas aplicadas – é menor, enquanto grupos de maior renda vivem em áreas de maior acessibilidade.

As métricas de acessibilidade vêm sendo explorada nas últimas décadas em diversas áreas, sobretudo no planejamento urbano e de transportes. Enquanto a acessibilidade cumulativa tem sido utilizada há algum tempo (Handy e Niemeier, 1997; Lopez, 2019), a BFCA é mais recente (Paéz, Higgins e Vivona, 2019; Desjardins, Higgins e Paéz, 2022). Revela-se importante também as análises de desigualdades espaciais de acessibilidade no território (Ferrari, Feitosa e Tomasiello, 2021; Giannotti et al, 2021).

2. Dados e Métricas

2. 1. Acessibilidade Cumulativa

Segundo Handy e Niemeier (1997), a acessibilidade cumulativa é considerada a classe mais simples de métricas de acessibilidade, uma vez que representa quantas oportunidades podem ser alcançadas dentro de um certo limite de tempo de viagem. A acessibilidade cumulativa retrata um leque de escolhas possíveis para os residentes acessarem, neste caso, as escolas localizadas em um raio de até 15 minutos de caminhada.

A equação da métrica cumulativa de acessibilidade aplicada neste estudo é a seguinte (Equação 1 baseada em Paéz; Scott; Morency, 2012):

$$A_{ik}^m = \sum_j Q_{jk} I(t_{ij}^m \leq L_i) \quad (\text{Equação 1})$$

Em que: A_{ik}^m é a acessibilidade cumulativa através do modo de transporte m (caminhada) às oportunidades do tipo k (escola) no local i ; Q_{jk} é a quantidade total de oportunidades do

tipo k no local j ; t_{ij}^m é o tempo de viagem entre o local i e o local j através do modo de transporte m ; L_i é o limite de tempo de viagem (15 minutos); I é um valor booleano, pois será igual a 1 se o tempo de viagem entre o local i e o local j através do modo de transporte m for menor ou igual ao limite de tempo de 15 minutos e será igual a 0 caso contrário.

Para o cômputo da métrica, são necessários pontos georreferenciados de origem das viagens, bem como os pontos de destino, além de uma rede de logradouros através da qual serão realizadas as viagens. Uma vez que a unidade espacial de análise foi a célula da grade estatística do Instituto Brasileiro de Geografia e Estatística (IBGE, 2016), os centroides de tais células foram considerados como os pontos de origem das viagens até as escolas. As células inseridas no perímetro urbano dos municípios possuem dimensões de 200 metros de altura e 200 metros de largura, enquanto no perímetro rural suas dimensões são 1 quilômetro de altura por 1 quilômetro de largura. Como pontos de destino das viagens foi utilizada a Base Cartográfica Digital Georreferenciada das Escolas da RMSP, disponibilizada pelo Centro de Estudos da Metrópole (CEM, 2016). A rede de logradouros foi, então, criada em um Sistema de Informações Geográficas, a partir da Base Cartográfica Digital Georreferenciada de Logradouros da RMSP (CEM, 2020).

A vantagem da métrica cumulativa de acessibilidade em relação às demais é a facilidade de interpretação de seus dados por especialistas de outras áreas do conhecimento. Entretanto, esta medida pondera todos os destinos potenciais que atendem ao limite de tempo de viagem de maneira igual, sem considerar o fator competição entre as oportunidades, diferente da métrica BFCA apresentada a seguir.

2. 2. Acessibilidade BFCA

A acessibilidade BFCA (Balanced Floating Catchment Areas) foi apresentada por Páez, Higgings e Vivona (2019), com o intuito de considerar a competição no acesso a oportunidades, porém sem provocar vieses de inflação de oferta e de demanda entre as oportunidades.

Neste estudo, a alocação de estudantes em relação aos professores em escolas dentro de um raio de caminhada do estudante de até 15 minutos foi analisada a partir desta métrica. Ou seja, foi possível compreender quantos potenciais alunos há para cada professor diante do limite de tempo de viagem estabelecido. Para tanto, o cômputo da acessibilidade BFCA foi realizado em três etapas, descritas a seguir.

A primeira etapa compreende a distribuição dos alunos em relação às escolas localizadas em, no máximo, 15 minutos de caminhada de distância da célula da grade estatística do IBGE onde o aluno reside (Equação 2).

$$P_k = \sum_{i=1}^n P_i w_{ik} \quad (\text{Equação 2})$$

Em que: P_k é a população alocada para a oportunidade k (escola); P_i é a população residente na célula i ; w_{ik} é a impedância que a população residente na célula i encontra para acessar a oportunidade k dentro do limite de viagem de 15 minutos de caminhada.

A segunda etapa consiste no cálculo do nível de serviço que cada escola comporta diante da quantidade de profissionais que possui e da quantidade de alunos que conseguem acessá-la dentro do limite de tempo estabelecido (Equação 3).

$$S_k = \frac{T_k}{P_k} \quad (\text{Equação 3})$$

Em que: S_k é o nível de serviço da oportunidade k (escola); T_k é o total de professores na oportunidade k ; P_k é a população alocada para a oportunidade k (primeira etapa).

Por fim, é possível computar a acessibilidade BFCA (A_i) na célula i , a partir de uma soma ponderada do nível de serviço das escolas acessadas pelos residentes da célula i dentro do limite de tempo de viagem determinado, considerando os pesos espaciais (Equação 4).

$$A_i = \sum_{k=1}^K S_k w_{ik} \quad \text{(Equação 4)}$$

Diante das duas métricas de acessibilidade aplicadas a este estudo, percebe-se que apesar de a BFCA, se comparada à acessibilidade cumulativa, requerer uma diversidade maior de dados, há um conjunto de dados comum entre elas. Assim, para além dos dados apresentados na seção 2.1., a BFCA requer o número de professores que trabalham em cada um dos equipamentos educacionais localizados na área de estudo e de quantos estudantes residem em cada célula da grade estatística. Tais informações foram obtidas a partir de dados do Censo Escolar de 2015, realizado pelo INEP, e Censo Demográfico de 2010, realizado pelo IBGE, e inseridos na escala da grade estatística do IBGE. Esta última informação foi calculada a partir de um filtro de idade aplicado à população de cada célula (considerou-se que crianças entre 6 e 17 anos frequentam escolas de ensino fundamental e médio). Os dados populacionais referentes ao número de professores e estudantes em cada célula representam certa limitação deste estudo, visto que há uma incompatibilidade espacial e um atraso temporal, já que o Censo Demográfico é do ano 2010 e o Escolar de 2015. Estes dados do Censo Escolar e Demográfico são dispostos em setores censitários, mas para fins do cômputo da BFCA foram agrupados em centroides e atribuídos por localização às células da grade estatística.

2. 3. Coeficiente de Correlação de Spearman

Para análise da relação entre os níveis de acessibilidade a escolas na RGABC e os grupos de renda, aplicou-se o coeficiente de correlação de Spearman, uma vez que é um índice não paramétrico ao não requerer que as variáveis envolvidas possuam distribuição normal bivariada (Sandoval, 2008). O coeficiente de Spearman varia entre -1 e 1, correspondendo a uma correlação negativa quanto mais próximo de -1 e a uma correlação positiva quando próximo de 1. Os grupos populacionais foram definidos de acordo com os dados de rendimento nominal mensal dos responsáveis por seu domicílio do Censo Demográfico de 2010 (IBGE).

3. Resultados e Discussão

A Figura 1 demonstra o resultado das métricas de acessibilidade Cumulativa e BFCA das células da grade estatística do IBGE às escolas, na RGABC. Nota-se, na Figura 1a, que as áreas em azul representam células com maior acessibilidade, uma vez que um estudante que caminha por até 15 minutos a partir dali pode acessar entre 4 e 9 escolas. Por outro lado, as áreas em vermelho coincidem com células onde um estudante ali residente não acessa escola alguma dentro deste limite de tempo de caminhada. Na Figura 1b, com relação à BFCA, por sua vez, percebe-se que as áreas em azul são espaços com bons índices de acessibilidade, em que cada professor de escola é responsável por até 12 crianças que caminham dentro do limite de viagem estabelecido, diferentemente das áreas em laranja, onde cada professor atende mais de 34 crianças. As áreas em vermelho da métrica BFCA, em contraste, demonstram áreas onde não há professor algum de escola em um raio de 15 minutos de caminhada para as crianças que residem nessas células.

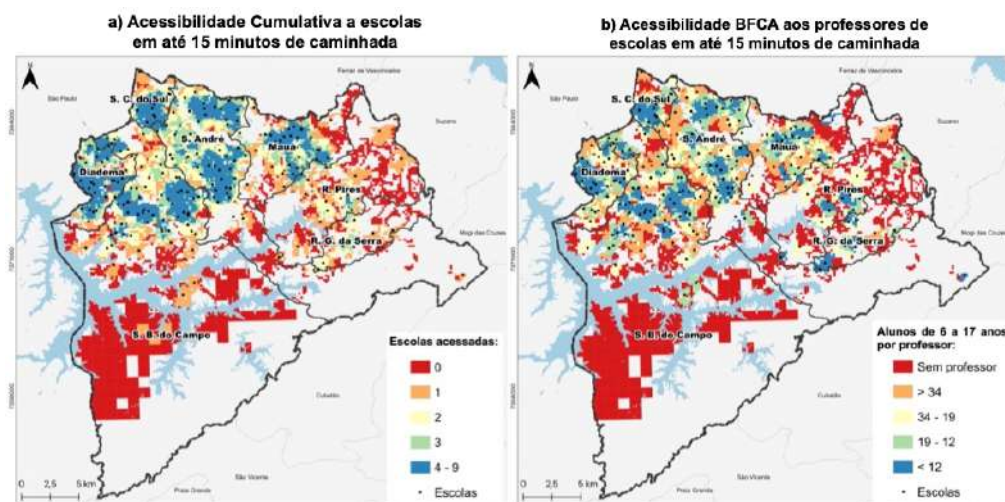


Figura 1. Métricas de acessibilidade a escolas de nível fundamental e médio em até 15 minutos de caminhada, na Região do Grande ABC, SP.

Ao comparar a distribuição espacial das duas métricas de acessibilidade a escolas de ensino fundamental e médio na RGABC, é possível observar que uma ressalta o padrão da outra, com sutis diferenças. No geral, destaca-se que as regiões centrais dos sete municípios apresentam acessibilidades a escolas melhores que suas demais áreas, guardadas as diferenças, entretanto, entre os sete centros. Nessa conjuntura, percebe-se que o município de São Caetano do Sul possui boa acessibilidade a escolas. Os municípios de Ribeirão Pires e Rio Grande da Serra revelaram-se como piores em relação às acessibilidades analisadas dentro da região de estudo.

Tabela 1. Correlação entre variáveis de acessibilidade a escolas e grupos de renda

Acessibilidade Cumulativa			BFCA		
Pessoas Responsáveis com Rendimento Nominal Mensal:	Coefficiente de Correlação de Spearman (rho)	p-valor	Pessoas Responsáveis com Rendimento Nominal Mensal:	Coefficiente de Correlação de Spearman (rho)	p-valor
Até 1 S.M.	-0,31094	<2,2E-16	Até 1 S.M.	-0,2954989	<2,2E-16
Entre 1 e 2 S.M.	-0,1416127	<2,2E-16	Entre 1 e 2 S.M.	-0,1546873	<2,2E-16
Entre 2 e 3 S.M.	0,1419232	<2,2E-16	Entre 2 e 3 S.M.	0,127457	3,95E-16
Entre 3 e 5 S.M.	0,3351969	<2,2E-16	Entre 3 e 5 S.M.	0,2988459	<2,2E-16
Mais de 5 S.M.	0,2432631	<2,2E-16	Mais de 5 S.M.	0,2406646	<2,2E-16
Sem rendimento nominal mensal	-0,1747518	<2,2E-16	Sem rendimento nominal mensal	-0,1503007	<2,2E-16

Diante do rendimento mensal das pessoas responsáveis por seu domicílio na RGABC, foi possível analisar a distribuição espacial de seis grupos de renda, tendo como referência o salário mínimo (SM). Para cada um desses grupos (até 1 SM, entre 1 e 2 SM, entre 2 e 3 SM, entre 3 e 5 SM, mais de 5 SM e sem rendimento) foi computada sua correlação com os índices de acessibilidade cumulativa e BFCA, a partir do coeficiente de correlação de Spearman (tabela 1). Atestou-se que os valores das correlações são significativamente diferentes de zero, uma vez que o p-valor de cada uma das correlações foi inferior a 0,05.

O teste de correlação entre os grupos de renda e as métricas de acessibilidade revelou, portanto, que os grupos de menor renda (até 1 SM e entre 2 e 3 SM) estão negativamente correlacionados às duas métricas, bem como o grupo sem rendimento mensal. Por outro lado, os grupos de maior renda (entre 2 e 3 SM, entre 3 e 5 SM e mais de 5 SM) estão positivamente correlacionados aos índices de acessibilidade. Isso ignifica que os grupos de menor renda ou sem renda tendem a residir em áreas onde a acessibilidade a escolas e a seus professores é menor, enquanto grupos de maior renda em áreas com escolas e professores mais acessíveis.

4. Considerações Finais

Diante deste estudo, contata-se que a acessibilidade é uma medida útil para a formulação de políticas públicas ao auxiliar na identificação de áreas e grupos populacionais com dificuldade de acesso a alguma oportunidade. Esse fator foi potencializado diante da adoção da grade estatística como unidade espacial frente à sua possibilidade de agregação de dados independentes de divisões político-administrativas, à estabilidade espacial ao longo do tempo (localização fixa ao longo do tempo) e ao seu alto grau de detalhamento espacial.

Especificamente em relação ao acesso às escolas, as métricas de acessibilidade calculadas no estudo auxiliam na decisão da administração pública sobre a distribuição espacial de equipamentos educacionais, bem como da concentração de professores neles.

Enquanto a acessibilidade cumulativa calculada ajuda na compreensão da demanda por escolas em cada área, a acessibilidade BFCA auxilia na compreensão da demanda por professores nessas áreas, bem como os recursos necessários. Ao passo que a métrica cumulativa possui uma interpretação mais simples, a BFCA é mais refinada ao considerar a demanda e competição entre oportunidades. Portanto, não é plausível considerar, no geral, uma medida melhor que a outra, uma vez que são complementares.

Referências Bibliográficas

- Centro de Estudos da Metrópole – CEM (2016). “Base Cartográfica Digital Georreferenciada das Escolas da RMSP”. São Paulo: CEM.
- Centro de Estudos da Metrópole – CEM (2020). “Base Cartográfica Digital Georreferenciada de Logradouros da Região Metropolitana de São Paulo”. São Paulo: CEM.
- Desjardins, E.; Higgins, C.; Paéz, A. (2022) “Examining equity in accessibility to bike share: A balanced floating catchment area approach”. Elsevier, v. 102
- Ferrari, T.; Feitosa, F.; Tomasiello, D. (2021) “Household structure and urban opportunities: Evaluating differences in the accessibility to jobs, education and leisure in São Paulo” *The Journal of Transport and Land Use*, v.14, n.1, p. 841-862
- Geurs, K.; Van Wee, B. (2004) “Accessibility evaluation of land-use and transport strategies: review and research directions”. *Journal of Transport Geography*, v. 12, n. 2, p. 127-140.
- Giannotti, M. et al (2021) “Inequalities in transit accessibility: Contributions from a comparative study between Global South and North metropolitan regions”. Elsevier, v. 109
- Handy, S.; Niemeier, D. (1997) “Measuring accessibility: an exploration of issues and alternatives”. *Environment and Planning*, v. 29, p. 1175-1194.
- Hansen, W. (1959) “How accessibility shapes land use. *Journal of the American Institute of Planners*”. v. 25, n. 2, p. 73-76.
- Instituto Brasileiro de Geografia e Estatística – IBGE. (2016) “Grade Estatística”. Rio de Janeiro: IBGE.
- Lopez, N.; Biona, J. (2019) “Assessing transport equity through a cumulative accessibility measure and Google Maps: a case study for healthcare in Metro Manila”. *A Companion to Transport, Space and Equity*, chapter 10, p. 136-146.
- Maricato, E. (2003) “Metrópole, legislação e desigualdade”. *Estudos Avançados*, v. 17, n. 4, p. 151-166.
- Paéz, A.; Higgins, C.; Vivona, S. (2019) “Demand and level of service inflation in Floating Catchment Area (FCA) methods”. *Plos One*, v. 14, n. 6.
- Paéz, A.; Scott, D.; Morency, C. (2012) “Measuring accessibility: positive and normative implementations of various accessibility indicators”. *Journal of Transport Geography*, v. 25, p. 141–153
- Sandoval, J. (2008) “Asociación entre variables: correlación no paramétrica”. *AMC*, v. 50, p. 144-146

Statistical Analysis of a Spatially Explicit Hydrological Forecasting Model using Remote Sensing of the Atmosphere

Jaqueline A. J. P. Soares^{1,2}, Michael M. Diniz³, Leonardo B. L. Santos^{1,2}

¹Center for Monitoring and Early Warning of Natural Disasters (Cemaden)
São José dos Campos – SP – Brazil

²National Institute for Space Research (INPE)
São José dos Campos – SP – Brazil

³Federal Institute of São Paulo (IFSP)
São José dos Campos – SP – Brazil

{jaqueline.soares,leonardo.santos}@cemaden.gov.br, michael.diniz@ifsp.gov.br

Abstract. *The performance analysis of a data based machine learning model depends directly on the separation of the data for training and testing, thus, a single performance evaluation can generate a biased results. Taking this into account, the aim of this paper is to propose and exemplify a statistical analysis method to get a proper understanding of a spatially explicit hydrological forecasting model performance. For this, the weather radar data used to training and testing a neural network based model was splitted randomically 100 times, and the RMSE and NASH metric was calculated for each one. The analysis of the metrics frequency distribution indicates that the neural network model had better performance than the persistence one.*

1. Introduction

According to [Mal et al. 2018], climate change has become an important factor for many environmental disasters in vulnerable communities and ecosystems. Disasters cause environmental impacts and abrupt changes in the population's daily life, causing financial and human losses.

Extreme hydrological events (EHEs), such as droughts and floods, vary spatially and temporally in nature, and their increase in the last few decades has motivated the research of the spatiotemporal variability of the future extreme precipitation and temperature [Diaz et al. 2019]. The distributed (or spatially explicit) hydrological models have increasingly contributed to these researches, as they consider the spatial variability of the study area, i.e., they use geoprocessing information to make the modeling more realistic.

During the creation, training and testing phases of a distributed hydrological model, there may be several stochastic processes involved. When dealing with stochastic processes, to provide a more adequate and reliable analysis, it is necessary to consider not just an isolated sample but, ideally, an ensemble of different samples. Otherwise, there is a risk that the analysis performed is biased and does not reflect reality.

This work presents a statistical analysis of a distributed hydrological forecasting model using remote sensing data from the atmosphere. This model was first introduced by [Freitas et al. 2022]. In this article, the performance of this model was compared with

Persistence, which will be more detailed in the Section 2.2. Furthermore, the stochastic processes present in the model were mapped and statistically analyzed. The results show that Persistence presented superior performance to the studied model for fast predictions of only 15 minutes. However, for very short-term forecasts, but with a slightly longer duration of 120 minutes, the model is already starting to show better results than Persistence (83% of the time).

2. Materials and Methods

This section presents the dataset used in this article, the sampling methodology for dividing the data between training and testing, the spatially explicit hydrological forecasting model, and the benchmark model used in the performance comparison.

2.1. Dataset

The study area covers the small watershed of the Bengalas river (Figure 1), which has approximately 190 km² of drainage area and is located in the mountainous region of Rio de Janeiro, Brazil. This region has a long history of disaster occurrences, mainly due to its natural characteristics and the large occupation of disaster risk areas.

The dataset used in this article is provided by the weather radar of Pico do Couto, in Petropolis, and operated by the Department of Airspace Control (DECEA). A weather radar emits high-energy electromagnetic waves to reach great distances. The analyzed data are from December 2011 to March 2013, with temporal resolution of 15 and 120 minutes. The reflectivity radar data (dBZ) was used to retrieve the accumulated volume of rainfall (mm) for 12 hours using the Marshall-Palmer relation [Marshall and Palmer 1948].

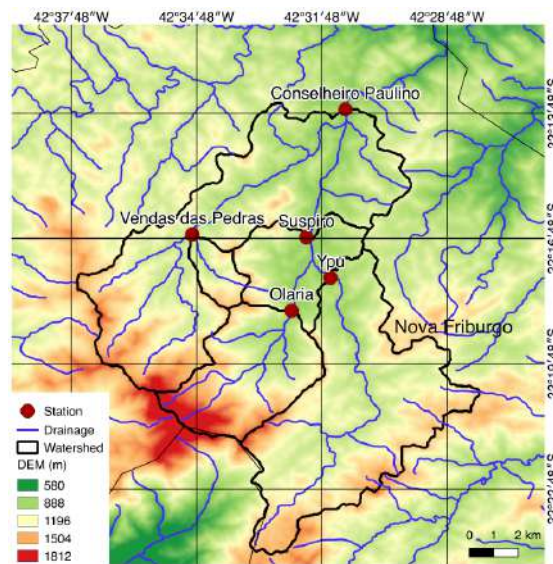


Figure 1. The area of study - Bengalas river watershed. Source: Adapted from [Freitas 2022].

In addition, the neural network training phase requires the water level data located at the lowest point of the watershed. This data was provided by the Conselheiro Paulino hydrological monitoring station of the State Environmental Institute (INEA), for the same period, with resolution of 15 minutes.

2.2. Neural Network and Persistence

The model used for the statistical tests is a neural network with Multilayer Perceptron (MLP) architecture, designed with three hidden layers and a single-neuron output layer. The neural network input is accumulated volume of rainfall for 12 hours derived from weather radar data, and the output is the water level data at the lowest point of the watershed. After training, the neural network used in this article is considered a deterministic model, i.e., it always produces the same output for the same set of inputs.

A basic performance analysis compares the neural network with a model that outputs as forecasting the replication of the last observed data (in this article, we name it "Persistence"). Consider $o(t)$ the observed level data for the time instant t and $y(t+1)$ the Persistence level prediction for the time instant $t+1$. Thus, we have that $y(t+1) = o(t) \forall t \geq 0$. The Section 3 presents the results of this comparison.

2.3. Training and Test Samples

Several stochastic processes may be involved during a neural network's creation, training and testing phases, such as initializing the neural network weights and dividing the dataset into training and test data. In this article, seeds were defined for all these processes to make it possible to reproduce the experiments. However, after creating the neural network, according to the mapping performed, the only stochastic process consists of dividing the data between training and testing.

In order to provide a more appropriate comparison of the performance of models based on the data used during the training and testing phase, it is recommended to analyze distributions of data splits between training and testing, and not just a single split. That is because if the data division is biased, even if random, it can result in performance metrics that do not reflect reality. For example, suppose a scenario where a given seed results in an excellent performance metric, but for most seeds, the result is poor. This scenario shows the importance of a broader assessment from a statistical point of view. Thus, this article presents a statistical analysis of this stochastic process to remove any obscurity in the use of the data.

Several data divisions (samples) were performed between the training and test set varying the seeds. In order to allow the reproduction of results, an initial seed (seed = 10) was used to generate another 100 random seeds, and a new data division between training and test was composed for each. It is important to note that the greater the number of seeds analyzed, the greater the reliability of the results, but the higher the total execution time required to produce the results.

Figure 2 shows examples of data divisions between training and testing, always keeping the proportion used of 80% of the instances for training, 10% for validation, and 10% for the neural network test.

The source code used in this article was developed in Python language. Regarding the division of data between training and testing, was used a method that splits arrays or matrices into random train and test subsets (*train_test_split* method from the *sklearn.model_selection* library [Scikit-learn.ORG 2022]). However, whenever randomization is part of a Scikit-learn algorithm, a *random_state* parameter may be provided to control the random number generator used. This work used this parameter to control the shuffling applied to the data before applying the split.

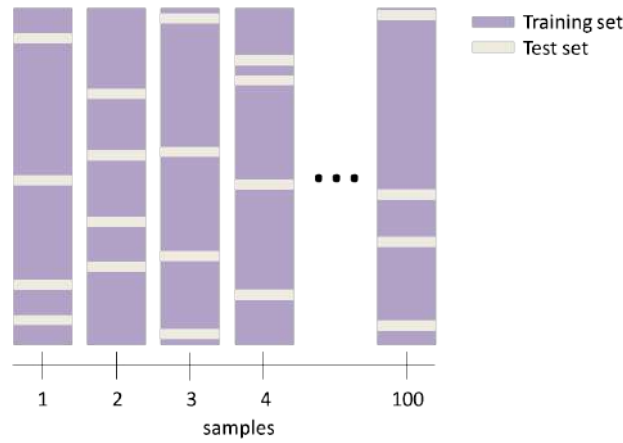


Figure 2. Examples of splitting data between the training and test set.

3. Results and Discussion

This section presents both the results obtained with the neural network against the ensemble of samples used in the stochastic process of training and testing, as well as the performance comparison between the neural network and Persistence.

Figure 3 shows the histograms of performance metrics *NASH* (a) and *RMSE* (b) of the neural network for 15-minute forecast, while Figure 4 shows histograms of the same metrics for 120-minute forecast. The dashed red line in the histograms represents Persistence's performance for the same metrics.

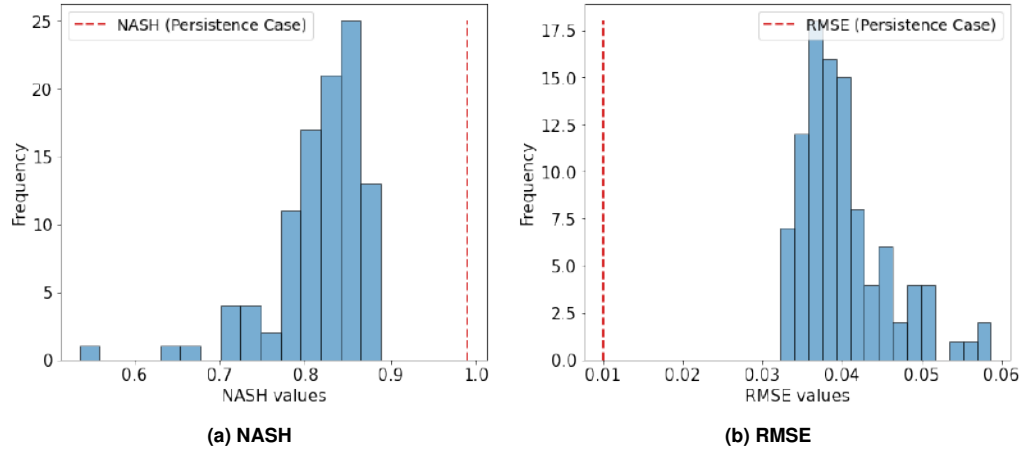


Figure 3. Distribution of the neural network metrics for 15-minute forecast.

Note that *NASH* and *RMSE* distributions are similar for the 15-minute (Figure 3) and 120-minute (Figure 4) forecasts, with better results for 120 minutes. Note that in Figures 3a and 4a, there is a greater concentration of samples in the highest values, with a respective median shift for this region of the graph and that this concentration quickly drops to lower values. On the other hand, in Figures 3b and 4b, there is a higher concentration of samples at lower values, and this concentration drops a little more gradually at higher values.

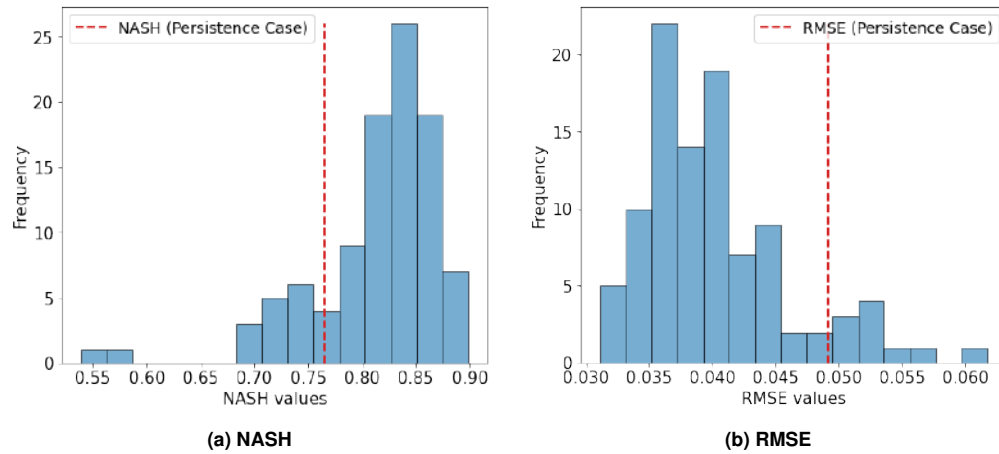


Figure 4. Distribution of the neural network metrics for 120-minute forecast.

Furthermore, it is possible to verify that Persistence surpassed the neural network for all the samples used for too short time intervals (Figure 3 - time in minutes). It is worth noting that the modeled problem is a spatial-temporal process with a not-so-fast time scale compared with the data's time resolution. However, for intervals of a few hours (Figure 4), the neural network presents more satisfactory results than Persistence - 83% of the seeds used in the data division between training and testing resulted in models with better values of *NASH* when compared to Persistence. In this case, there are significant changes in the spatial-temporal dynamics, and Persistence is not so far good as previously.

Table 1 shows a summary of the neural network performance metrics statistics for the 15 and 120-minute forecasts against the 100 samples used in the stochastic process of dividing the data between training and test set.

Table 1. Statistical indicators of neural network performance metrics for 15-minute and 120-minute forecasts.

Statistical Indicators	15 minutes		120 minutes	
	RMSE	NASH	RMSE	NASH
mean	0.0403	0.8170	0.0403	0.8122
std	0.0055	0.0554	0.0057	0.0613
min	0.0322	0.5363	0.0310	0.5389
25%	0.0362	0.7978	0.0363	0.7907
50%	0.0391	0.8326	0.0392	0.8273
75%	0.0426	0.8558	0.0430	0.8516
max	0.0587	0.8886	0.0617	0.8988

According to Table 1, statistics for metrics *RMSE* and *NASH* were quite similar between the 15 and 120-minute forecasts. Furthermore, for 120 minutes, the minimum *NASH* was 0.5389, and the maximum was 0.8988, resulting in a range of 0.3599. This variation is very significant for this metric. The same is not valid for *RMSE*, whose minimum and maximum values were, respectively, 0.0310 and 0.0617, resulting in a range of 0.0307, which is not very representative for this metric.

4. Conclusions and Perspectives

In this work, we propose and exemplify a statistical analysis method to properly understand a spatially explicit hydrological forecasting neural network using weather radar data. In this method, the stochastic process of dataset separation between training and test was analyzed not as an isolated sample but as an ensemble of different samples.

The neural network's performance has been compared with Persistence, whose prediction consists of replicating the last observed data. For a 15-minute (very short-term) forecast scenario, Persistence outperformed the neural network in 100% of the samples, and for 120 minutes (short-term), the neural network outperformed Persistence in 83% of the samples. This last case shows that the choice of seed (sample) can be decisive in the performance comparison, which justifies and motivates the analysis of different samples.

Limitations of this work are the extension of the time series used in separating data between training and test and the number of analyzed seeds. Currently, the time series is limited to a few years of data. However, as future work, we intend to increase it to allow a better analysis of the stochastic process analyzed in this article. Furthermore, we intend to increase the number of seeds by an order of magnitude to increase the results' reliability further.

5. Acknowledgments

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

References

- Diaz, V., Corzo, G., and Pérez, J. R. (2019). 3 - Large-scale exploratory analysis of the spatiotemporal distribution of climate projections: Applying the striving toolbox. In Corzo, G. and Varouchakis, E. A., editors, *Spatiotemporal Analysis of Extreme Hydrological Events*, pages 59–76. Elsevier.
- Freitas, C. P. (2022). Neural network for very short-term hydrological forecasting. Master's Dissertation in Applied Computing, National Institute for Space Research (INPE), São José dos Campos.
- Freitas, C. P., Bacelar, L., Soares, J. A. J. P., Diniz, M. M., Lima, G. R. T., Stephany, S., and Santos, L. B. L. (2022). A neural network-based short-term hydrological model using weather radar data. *to be submitted*.
- Mal, S., Singh, R. B., Huggel, C., and Grover, A. (2018). *Introducing Linkages Between Climate Change, Extreme Events, and Disaster Risk Reduction*, pages 1–14. Springer International Publishing, Cham.
- Marshall, J. S. and Palmer, W. M. K. (1948). The distribution of raindrops with size. *Journal of the Atmospheric Sciences*, 5(4):165 – 166.
- Scikit-learn.ORG (2022). API Reference. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html (Accessed: 09 May 2022).

Deforestation Increases Within Full Protected Areas of the Brazilian Amazon

Guilherme Mataveli¹, Alber Sanchez¹, Michel Chaves¹, João Guerrero²

¹Divisão de Observação da Terra e Geoinformática – Instituto Nacional de Pesquisas Espaciais (INPE)
Avenida dos Astronautas, 1758, Jardim da Granja – 12227-010 – São José dos Campos – SP – Brasil

²Centro Nacional de Pesquisa em Pesca, Aquicultura e Sistemas Agrícolas – Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA)
Avenida NS10, Loteamento Água Fria – 77008-900 – Palmas, TO, Brasil
{guilherme.mataveli, alber.ipia, michel.chaves}@inpe.br,
jvguerrero2@gmail.com

Abstract. *Since 2012, deforestation rates have been increasing within the Brazilian Amazon. This situation undermines the mitigation efforts against global climate change and stresses our need for both accurate identification of deforestation hotspots and promotion of law enforcement. It also implies a permanent need to explore the deforestation data, looking for patterns that enable the public, the academia, and authorities to anticipate and prepare for deforestation outburst. For this reason, we analyzed data of the Brazilian deforestation monitoring program (PRODES) to find deforestation patterns within Full Protected Areas of the Brazilian Amazon. Our aim is to identify and quantify deforestation from 2008 to 2021. We applied a trend analysis to find FPAs with significant increments and we found that approximately 15% of FPAs have statistically significant deforestation growth according to the Mann-Kendall test at 5% significance. Furthermore, we found that only four FPAs accounted for 67.6% of the deforestation increments during 2021, making them good candidates for closer monitoring and law enforcement. Besides, our results also show the role of FPAs as forest safeguards, although deforestation within them is increasing.*

1. Introduction

Deforestation is a major global anthropogenic disturbance [Seymour and Harris 2019; Wade et al 2020], even when intact forests offer important ecosystem services [Watson et al. 2018] such as climate regulation. Therefore, the preservation of natural undisturbed forests is of utmost importance for human well-being. The declaration of Protected Areas (PAs) is the most effective tool for reducing deforestation since deforestation rates are lower within PAs than in neighbor unprotected areas [Johnson et al. 2017; Fuller et al. 2020]. However, even while PAs are crucial for protecting standing forests, many of them remain vulnerable to illegal activities, such as deforestation and mining, due to their dependency on government and law enforcement [Nolte et al. 2013].

The Brazilian Amazon (BA) is an example of this situation. The BA has been facing a deforestation resurgence since 2019, mostly associated with the weakening of environmental laws occurring in Brazil. Deforestation rates over the last three years have surpassed 10,000 km² in the BA [INPE, 2022]. After a historical deforestation reduction of 84% in 2012 compared to 2004, the deforestation rate in 2021 has reached the highest value during the last 15 years [INPE, 2022]. Unfortunately, this recent deforestation increase is not restrained to unprotected areas but is also occurring within PAs: (i) the annual average deforestation rate within Indigenous Lands (ILs) of the BA in 2021 was 52% above the average annual rate from 2008 to 2021, and (ii) the deforestation rate within Conservation Units (CoUs) of the BA in 2021 was the highest since 2008 [INPE 2022].

When referring to CoUs, they are classified as Environmental Protection Areas (EPAs), where occupation is allowed with settlers having to follow specific regulations, or Full Protection Areas (FPAs), where human occupation is banned. The increase in deforestation within FPAs is the most concerning threat to the preservation of BA standing forests. Due to the extension of the FPAs within the BA (more than 430 thousand km²), orbital remote sensing is the most cost-effective way to monitor deforestation within them all. In this study, we have used the openly available PRODES dataset [INPE 2022], the Brazilian official deforestation monitoring data, to identify and quantify, temporally and spatially, the increase in deforestation within FPAs of the BA from 2008 to 2021. Moreover, we have also identified the FPAs where deforestation is concentrated in the current period (the year of 2021), which are the ones where legal authorities must focus on the counteraction measurements to curb deforestation. A trend analysis was also performed to identify in which FPAs of the BA the increase in deforestation was statistically significant during the 2008-2021 time series.

2. Materials and Methods

2.1. Full Protection Areas

FPAs are the most protected type of CoUs in Brazil. They act as a barrier against human disturbances to the standing forests. The constant undermining of the Brazilian environmental policy has been linked to the increasing deforestation within FPAs because it encourages and legalizes illegal activities, reduces environmental fines, and decreases law enforcement actions [Artaxo 2019; Conceição et al. 2021].

The BA includes 95 FPAs, accounting for approximately 10% of its total area (431,448 km²). FPAs are classified into 5 subcategories: ecological station, park, natural monument, biological reserve, and wildlife refuge. Parks are the predominant FPAs in the BA. Although the increasing anthropogenic pressure over FPAs, they are primarily covered by forest formations [MapBiomass 2022].

2.2. Deforestation Data

The analysis of historical deforestation patterns was performed using the national official deforestation monitoring dataset, the Brazilian Amazon Deforestation Monitoring Program (PRODES). PRODES monitors clear-cut deforestation and provides annual official deforestation rates since 1988 using medium spatial resolution orbital images. The annual estimates of deforestation are based on visual interpretation

of the satellite images. Therefore, it is a reliable dataset to monitor deforestation increases in the BA [Almeida et al. 2016].

PRODES produces deforestation rates corresponding to the period from August 1st to July 31st of the following year. For example, PRODES 2021 was estimated considering deforestation from August 1st, 2020, to July 31st, 2021. An area mapped as deforested by PRODES is not analyzed in the following years, hence the deforestation rate is defined as increment in relation to previous years. The validation of PRODES estimates reached an overall accuracy of 93% [Almeida et al. 2016].

We downloaded the PRODES annual deforestation increments for the 2008-2021 period from the TerraBrasilis platform (<http://terrabrasilis.dpi.inpe.br/downloads/>) and then clipped the deforestation polygons to the delimitation of the FPAs made available by the Brazilian Ministry of Environment. At the end of this step, we had annual shapefiles quantifying the annual deforestation within the 95 FPAs of the BA.

2.3. Detecting Trends in Deforestation Within Full Protection Areas

Following the 2008-2021 time series, we have identified trends in deforestation of the FPAs within the BA. To this end, we have applied the non-parametric Mann-Kendall trend test, which determines if a trend exists in a time series [Cavalcante et al. 2019; Carvalho et al. 2020]. The Mann-Kendall test was applied to the 62 FPAs of the BA where deforestation is present in at least one year of the time series. Our null hypothesis is that there is no trend in the deforestation time series, while our alternative hypothesis is that there is. If the p-value is lower than the significance level of 5% the null hypothesis is rejected, proving that there is a statistically significant deforestation trend in the time series.

3. Results and Discussions

Figure 1 shows the annual deforestation increments within FPAs of the BA from 2008 to 2021 estimated by PRODES. We observe a reduction in deforestation starting in 2008 and ending in 2014, when the lowest estimate was produced (16.6 km²). This pattern follows the same one identified for the BA as a whole, explained by the Action Plan for Prevention and Control of Deforestation (PPCDAm) which promoted a severe conservation reform that decreased deforestation specially from 2007 to 2012 [West and Fearnside 2021].

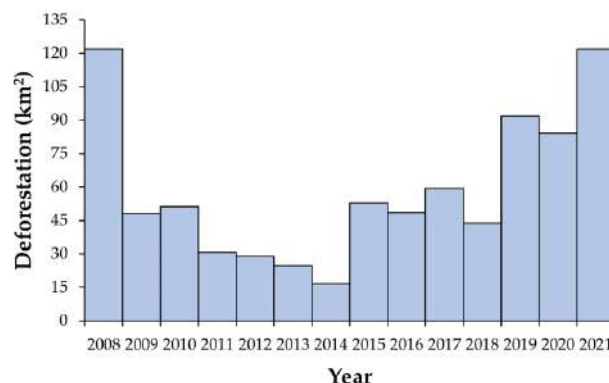


Figure 1. Annual deforestation increments within Full Protection Areas of the Brazilian Amazon during the 2008-2021 period.

Starting in 2015, deforestation increments within FPAs of the BA increased again and reached its highest value in the PRODES digital era in 2021 (121.9 km², 634% higher than the lowest estimate found in 2014). The observed increase resulted of several actions taken over the past years that indirectly incentivized land grabbing, such as the possible decriminalization of illegal occupation on public lands and the interruption of the PPCDam. In 2021, FPAs accounted for 8.6% of the deforestation increment in CoUs of the BA. Still, four FPAs accounted for 67.6% of the deforestation increment in these protected areas in 2021 and therefore are the ones requiring closer monitoring and law enforcement. These are the Terra do Meio Ecological Station (29.5 km²), the Guajará-Mirim State Park (28.9 km²), the Jamaxim National Park (13.4 km²), and the Nascentes Serra do Cachimbo Biological Reserve (10.7 km²). Three of them are inserted in the center of the “Arc of Deforestation” at the Pará State, while the Guajará-Mirim State Park is located in Rondônia State (Figure 2).

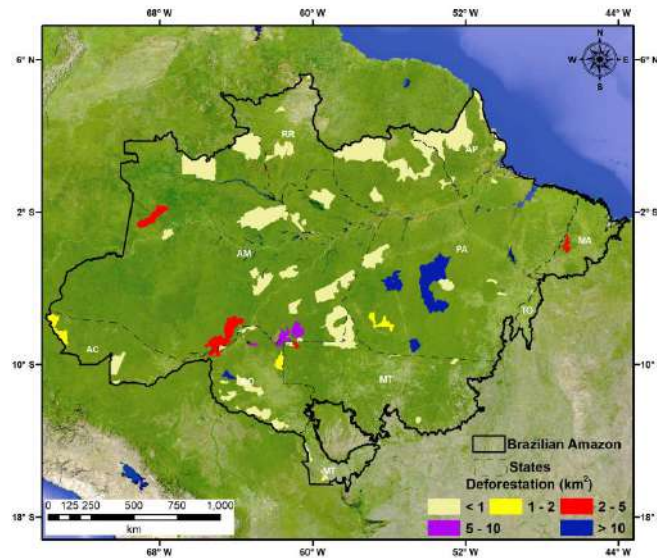


Figure 2. Deforestation increments within Full Protection Areas of the Brazilian Amazon in 2021.

An analysis of Figure 3 shows that 9 of the 62 FPAs of the BA where deforestation was detected in at least one of 14 years analyzed (~15% of them) had a statistically significant increasing trend according to the Mann-Kendall test at 5%. These were located not only at the “Arc of Deforestation”, but also in other areas, such as the State of Amapá, which may potentially constitute new deforestation hotspots. Deforestation in these areas is critical and must be counteracted now.

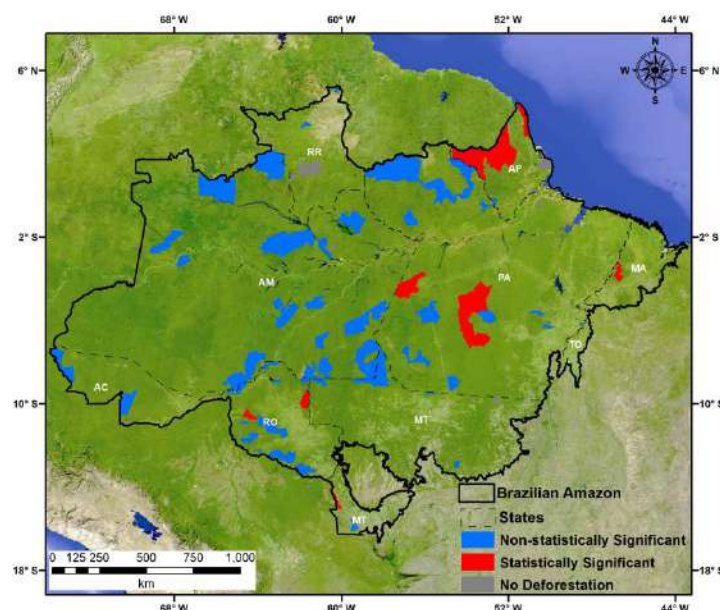


Figure 3. Mann–Kendall test applied to the deforestation time series of the Full Protection Areas within the Brazilian Amazon during 2008-2021.

4. Conclusions

In this work, we have analyzed the increase of deforestation within FPAs of the BA during the 2008-2021 period based on PRODES estimates. In this initial analysis, we observed that despite the recent push to ease environmental regulations in Brazil, FPAs are more effective at avoiding deforestation than unprotected areas. Therefore, law enforcement in these areas is key for maintaining BA standing forests and to stop new deforestation hotspots from appearing. However, we must emphasize that political calculus or lack of a proper monitoring can cancel conservation efforts within FPAs. Finally, PRODES is proven, once again, as a critical tool for legal authorities in curbing deforestation.

Acknowledgements

We thank FAPESP (grants 2019/25701-8 and 2021/07382-2) and CNPq (grants 380656/2022-1 and 350820/2022-8) for financial support.

References

- Almeida, C., Coutinho, A. D. A. C., Esquerdo, J. C. D. M., Adami, M., Venturieri A., Diniz, C. G., Dessay, C. G., Durieux, L. and A. R. Gomes. (2016). “High Spatial Resolution Land Use and Land Cover Mapping of the Brazilian Legal Amazon in 2008 Using Landsat-5/TM and MODIS Data”. *Acta Amazonica* 46 (3): 291–302. doi:10.1590/1809-4392201505504.
- Artaxo, P. (2019). “Working Together for Amazonia”. *Science* 363 (6425): 323. doi:10.1126/science.aaw6986.

- Carvalho, S., Oliveira, A., Pedersen, J. S., Manhice, H., Lisboa, F., Norguet, J., de Wit, F. and Santos, F. D. (2020) “A changing Amazon rainforest: Historical trends and future projections under post-Paris climate scenarios”. *Glob. Planet. Change*, 195: 103328. doi: 10.1016/j.gloplacha.2020.103328.
- Cavalcante, R. B. L., Pontes, P. R. M., Souza-Filho, P. W. M., and Souza, E. B. (2019) “Opposite Effects of Climate and Land Use Changes on the Annual Water Balance in the Amazon Arc of Deforestation”. *Water Resources. Research*, 55: 3092–3106. doi: 10.1029/2019WR025083.
- Conceição, K. V., Chaves, M. E. D., Picoli, M. C. A., Sánchez, A. H., Soares, A. R., Mataveli, G. A. V., Silva, D. E., Costa, J. S., and Camara, G. (2021) “Government policies endanger the indigenous peoples of the Brazilian Amazon”. *Land Use Policy*, 108: 105663. doi: 10.1016/j.landusepol.2021.105663.
- Fuller, C., Onde, S., Brook, B. W., and Buettel, J. C. (2020). “Protected-area planning in the Brazilian Amazon should prioritize additionality and permanence, not leakage mitigation”. *Biol. Conservation*, 248: 108673. doi: 10.1016/j.biocon.2020.108673.
- INPE (National Institute for Space Research). (2022). “Deforestation monitoring of the Brazilian Amazon rainforest and Cerrado biome by satellite”. https://terrabrasilis.dpi.inpe.br/app/dashboard/deforestation/biomes/legal_amazon/incrments
- Johnson, C. N., Balmford, A., Brook, B. W., Buettel, J. C., Galetti, M., Guangchun, L., J. M., and Wilmschurst, J. M. (2017). “Biodiversity Losses and Conservation Responses in the Anthropocene”. *Science* 356 (6335): 270–275. doi:10.1126/science.aam9317.
- MapBiomas. (2022). “Collection 6.0 of Brazilian Land Cover & Use Map Series”. Available online: <http://mapbiomas.org/en>.
- Nolte, C., Agrawal, A., Silvius, K. M., and Soares-Filho, B. S. (2013). “Governance regime and location influence avoided deforestation success of protected areas in the Brazilian Amazon”. *PNAS*, 110 (13): 4956–4961. doi: 10.1073/pnas.1214786110.
- Seymour, F., and Harris, N. L. (2019). “Reducing Tropical Deforestation”. *Science* 365 (6455): 756–757. doi:10.1126/science.aax8546.
- Wade, C. M., Austin, K. G., Cajka, J., Lapidus, D., Everett, K. H., Galperin, D., Maynard, R., and Sobel, A. (2020). “What Is Threatening Forests in Protected Areas? A Global Assessment of Deforestation in Protected Areas, 2001–2018”. *Forests* 11 (5): 539. doi:10.3390/f11050539.
- Watson, J. E. M., Evans, T., Venter, O., Williams, B., Tulloch, A., Stewart, C., Thompson, I., et al. (2018). “The Exceptional Value of Intact Forest Ecosystems”. *Nature Ecology and Evolution* 2 (4): 599–610. doi:10.1038/s41559-018-0490-x.
- West, T. A. P., and Fearnside, P. M. (2021). “Brazil’s conservation reform and the reduction of deforestation in Amazonia”. *Land Use Policy*, 100, 105072. doi: 10.1016/j.landusepol.2020.105072.

Análise da cobertura da terra no Parque Nacional Yasuní, Equador, usando Random Forest e computação em nuvem

Andrés Velastegui-Montoya^{1,2}, Hugo Rivera-Torres¹, Viviana Herrera-Matamoros¹, Luis sadeck³, Renata Pacheco Quevedo⁴

¹Facultad de Ingeniería en Ciencias de la Tierra (FICT) - ESPOL Polytechnic University
- Guayaquil, Guayas - Ecuador

²Centro de Investigación y Proyectos Aplicados a las Ciencias de la Tierra (CIPAT) -
ESPOL Polytechnic University - Guayaquil, Guayas – Ecuador

³Doutorando do Programa de Pós-Graduação em Ciências Ambientais – Universidade
Federal do Pará (UFPA) - Belém, PA - Brasil.

⁴Divisão de Observação da Terra e Geoinformática - Instituto Nacional de Pesquisas
Espaciais (INPE) – São José dos Campos, SP - Brasil

{dvelaste, hrrivera, isauhm}@espol.edu.ec

luis.sadeck@ufpa.br

renata.quevedo@inpe.br

Abstract. *The Yasuní National Park (YNP) is an Ecuadorian protected area recognized for its biodiversity. However, in recent decades, human activities have altered its forest cover. The present study proposes the classification of the YNP land cover during the year 2019. This paper is structured in three phases. The first phase focuses on data collection and preparation. The second phase applies land cover classification using the geoprocessing tool of the Google Earth Engine cloud. The third phase corresponds to the post-processing classification using the ArcGIS Pro software. The results revealed the presence of human activities in the study area.*

Resumo. *O Parque Nacional Yasuní (PNY) é uma área equatoriana protegida, reconhecida por sua biodiversidade. Nas últimas décadas, as atividades antropogênicas alteraram a cobertura florestal da área. Assim, este estudo propôs classificar a cobertura da terra no PNY para o ano 2019. O estudo foi processado em três fases. A primeira fase corresponde à coleta e preparação de dados. A segunda consistiu na classificação da cobertura da terra, usando geoprocessamento na nuvem do Google Earth Engine. A terceira fase correspondeu ao pós-processamento da classificação no software ArcGIS Pro. Dessa maneira, os resultados revelaram a presença de atividades antropogênicas no PNY.*

1. Introdução

O Parque Nacional Yasuní (PNY) é uma área protegida, localizada na Amazônia equatoriana, reconhecido por sua biodiversidade [Finer et al. 2009]. O PNY foi criado no ano 1979, concomitantemente à delimitação do território da comunidade indígena Huaorani. A área do parque divide-se em quatro zonas: i) área protegida do PNY; ii) área

de proteção integral de Tagaeri-Taromenane; iii) área de amortecimento; iv) zona de exploração de petróleo [Finer et al. 2009; Ministerio del Ambiente 2011]. Em 1989, o PNY foi declarado como reserva mundial da biosfera pela UNESCO, o que restringe a realização de atividades extrativas [Bass et al. 2010].

Entretanto, o território do PNY e seu entorno foram afetados pelo *boom* do petróleo, acarretando uma série de mudanças [Finer et al. 2009; Thieme et al. 2018]. Primeiramente, a indústria do petróleo causou a redução da área florestal, devido à construção de infraestruturas e perfurações de poços de petróleo [Llerena-Montoya et al. 2021; Sellers et al. 2017]. Da mesma forma, estradas de acesso foram construídas para facilitar o transporte de trabalhadores e equipamentos, além de instalação de oleodutos para o escoamento de petróleo e seus derivados [Finer et al. 2009]. Posteriormente, assentamentos humanos foram formados ao longo das estradas, o que impulsionou a criação de áreas agropastoris na região [Gray and Bilsborrow 2020]. Essas atividades produziram problemas relacionados à poluição petrolífera e à invasão da área protegida do PNY [Kingsbury et al. 2019].

Assim, é necessário monitorar a redução das áreas florestais e otimizar o uso da terra, com a intenção de mitigar a crise climática e ecológica [Bass et al. 2010; Lovejoy and Nobre 2019]. Nesse contexto, este trabalho tem como objetivo conhecer o estado de uso e cobertura da terra no PNY no ano 2019, a fim de identificar o impacto antropogênico nas áreas protegidas, através do mapeamento de uso e cobertura da terra na região. Para isso, utilizamos o algoritmo de aprendizado de máquina *Random Forest* e executamos o processamento em nuvem, através da plataforma *Google Earth Engine* (GEE). O *Random Forest* se destaca por ser um algoritmo de *ensemble learning*, que considera a média das previsões de todas as Árvores de Decisão geradas, tendendo a apresentar menores erros. Por outro lado, o GEE reduz o tempo de processamento das imagens e permite realizar modificações rapidamente, sem a necessidade de repetir todo o processo [Parente et al. 2019; Tamiminia et al. 2020].

2. Metodologia

O PNY está localizado na região da Amazônia equatoriana (0°26'32"S, 75°25'02"O) e (01°42'12"S, 76°46'04"O), entre as províncias de Orellana e Pastaza (Figura 1), com 1.022.736 ha de extensão [Finer et al. 2009; Ministerio del Ambiente 2011]. A região apresenta um clima tropical úmido, de acordo com a classificação Köppen [Peel et al. 2007], com chuvas constantes ao longo do ano e precipitação média anual de 3.200 mm [Olivares et al. 2017]. Sua rica biodiversidade é composta por cerca de 1.138 espécies de fauna que habitam a floresta úmida do parque, a qual é composta por cerca de 2.274 espécies de árvores e arbustos [Ministerio del Ambiente 2011].

A aquisição dos dados foi feita a partir da coleção de imagens Sentinel-2 Level-1C (reflectância no topo da atmosfera), disponíveis na plataforma GEE. A partir das imagens disponíveis entre 1 de janeiro e 31 de dezembro de 2019, foram selecionadas imagens com cobertura de nuvens menor do que 20% para geração de um mosaico representativo do cenário de 2019. Consideraram-se as bandas X, Y, Z para a classificação do uso e cobertura da terra. A banda QA60 foi utilizada para aplicação de uma máscara de nuvens, em que os *pixels* com presença de nuvens receberam valor nulo. Para ordenar cada pilha de *pixels* e formar o mosaico, foi computada a mediana de cada *pixel*, utilizando-se os valores não-nulos. A área do PNY foi importada do banco de dados *The World Database on Protected Areas* (WDPA) em formato vetorial [UN Environment

Programme World Conservation Monitoring Centre 2021].

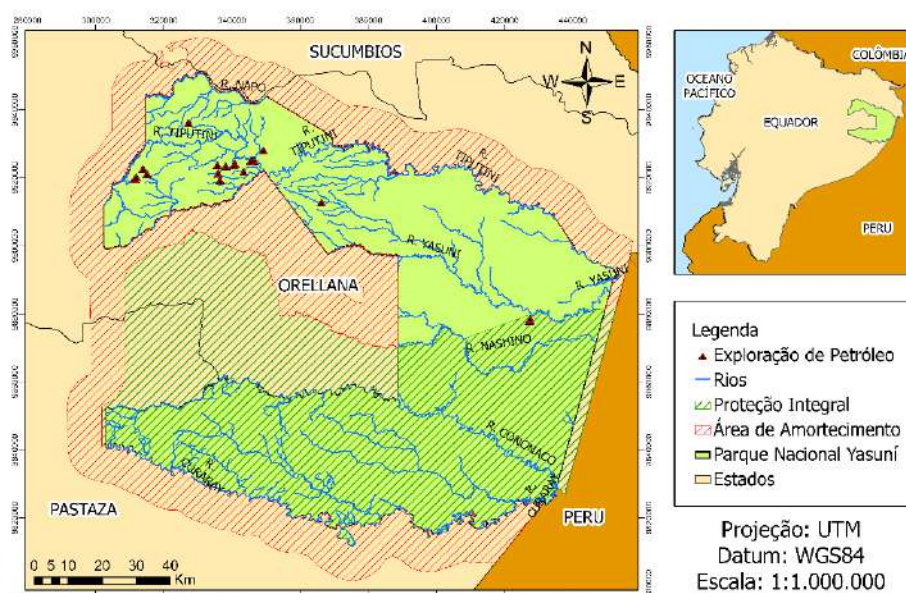


Figura 1. Localização do Parque Nacional Yasuní, Equador.

A classificação do uso e cobertura da terra foi realizada segundo as respostas espectrais dos alvos, considerando-se as classes: 1) solo exposto: composto por planícies aluviais e bancos de areia; 2) área antropizada: inclui as modificações florestais decorrentes de atividades humanas como desflorestamento e degradações, que posteriormente foram convertidos em agricultura ou pecuária; 3) floresta: cobertura florestal sem intervenção humana; 4) floresta sob nuvem de baixa densidade: todos os tipos de floresta coberta por nuvens de baixa densidade; 5) água: representada por lagos; 6) rios: correspondente a cursos d'água, identificados como marrom claro na imagem RGB; 7) nuvem: áreas sem informação espectral devido à cobertura de nuvem; 8) sombra: áreas sem informação espectral devido à cobertura de sombra de nuvem; 9) área urbana: inclui a expansão urbana e a infraestrutura gerada para extração de petróleo.

Sendo assim, foi desenvolvido um código em *JavaScript* para classificação no GEE. O classificador utilizado foi o Random Forest, algoritmo de aprendizado de máquina que cria e combina aleatoriamente Árvore de Decisão para gerar os resultados da classificação para cada pixel [Breiman 2001; Thonfeld et al. 2020]. Entre os parâmetros a serem definidos para o classificador, foram consideradas 30 árvores.

Finalmente, o raster classificado no GEE foi importado para o software *ArcGIS Pro*, onde se fez agrupamento de classes, resultando em seis classes de cobertura da terra. Dessa forma, o uso e cobertura da terra foi reclassificado em: 1) floresta: floresta e floresta sob nuvem de baixa densidade; 2) água: lagos e rios; 3) solo exposto; 4) área urbana; 5) área antropizada; e 6) área não identificada: sombra e nuvem. Além disso, foi feito o cálculo das áreas de cada classe.

3. Resultados e Discussões

O algoritmo Random Forest, utilizado para a classificação supervisionada, apresentou uma precisão de 99,97%. O resultado da classificação da cobertura da terra é apresentado

na Figura 2. A classificação LULC dentro do YNP está disponível em: <https://code.earthengine.google.com/90a86f18a518c3beabf239bf93dd9e87>. Da mesma forma, a Tabela 1 mostra as áreas em hectares e a porcentagem relativa ocupada por cada uma das classes (Tabela 1).

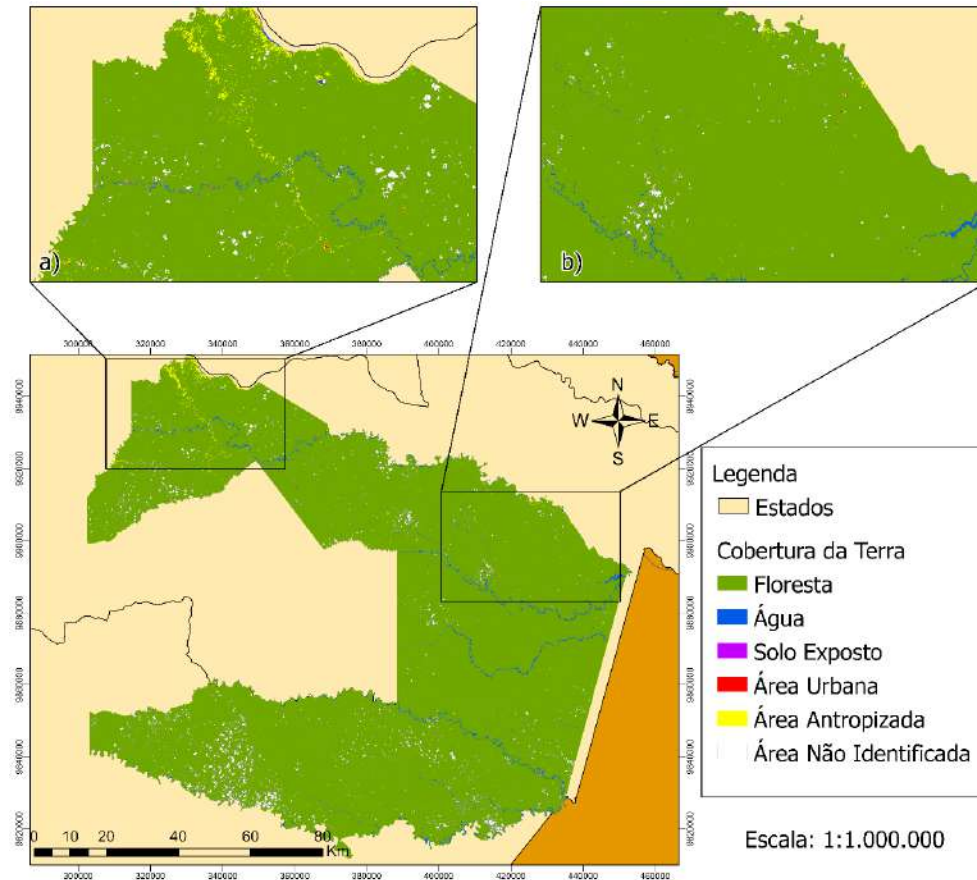


Figura 2. Classificação da cobertura da terra no Parque Nacional Yasuní em 2019. Áreas com maiores mudanças de cobertura da terra (a) e (b).

Tabela 1. Quantificação das áreas das classes de cobertura da terra no YNP

Classe	Área [ha]	Área [%]
Floresta	994.829,8	97,231
Área antropizada	736,1	0,072
Solo exposto	1.267,4	0,124
Área urbana	43,4	0,004
Água	4.345,8	0,425
Área não identificada	21.942,4	2,144

Como resultado do mapeamento da área de estudo, pode-se verificar que as áreas urbanas e antropizadas representaram 0,08% do PNY. Estas áreas estão localizadas na zona norte do parque, setor que se localizam as estradas de acesso às instalações petroleiras, segundo as Figuras 2a e 2b. As áreas antropizadas abrangeram 736,09 ha, o que representa 0,07% da área total do PNY. Essas áreas estão localizadas no entorno das

cidades e de áreas com solo exposto (Figura 2a), apresentam interesse agropecuário e têm substituído a cobertura florestal. A presença de manchas urbanas também revela a atividade humana. Estas áreas cobriram 43,35 ha, menos de 0,01% da área de estudo. As áreas urbanas estão localizadas nas imediações das áreas de solo exposto associadas às vias de acesso. A principal delas é a rodovia Maxus, que liga as áreas urbanas aos poços de extração de petróleo. Além disso, foi possível identificar áreas urbanas no entorno de corpos d'água, vinculada às populações e comunidades indígenas da região.

O solo exposto cobriu uma área de 1267,37 ha, o que equivale ao 0,12% da área de estudo. Esta classe corresponde a vias de acesso, solo exposto e bancos de areia ao redor das áreas urbanas e nas margens dos rios. Os corpos d'água cobriram 4.345,83 ha, equivalente a 0,42% do PNY. Os rios Conocaco, Nashiño e Tiputini atravessam de oeste para leste o PNY, enquanto o rio Curaray delimita o sul do PNY (Figura 1).

A classe predominante foi floresta, a qual abrangeu 994.829,80 ha, equivalente a 97,23% do PNY. Essas extensões de áreas florestais correspondem à floresta tropical úmida de terra firme. Além disso, em 2019, mesmo havendo atividades antrópicas no PNY, não foram identificadas atividades antrópicas na Área de proteção integral de Tagaeri-Taromenane, dado que as atividades extrativistas e caça estão proibidas nesta parte do parque [Finer et al. 2009; Ministerio del Ambiente 2011]. Finalmente, a classe de áreas não identificadas indicou que não foram obtidas informações espectrais para 21.942,39 ha (2,14%). Isso devido à forte presença de nuvens na região durante a maior parte do ano.

Dessa forma, a classificação de uso e cobertura da terra mostrou como os trabalhos de engenharia no PNY influenciaram direta e indiretamente na cobertura florestal [Thieme et al. 2018]. A construção de rodovias e vias de acesso tem gerado mudanças prejudiciais à manutenção do PNY e do seu entorno, como a erosão do solo, estabelecimento de assentamentos e atividades agrícolas.

4. Conclusões

O mapa de uso e cobertura da terra revelou a presença de atividades antrópicas na região. Da mesma forma, foi possível identificar a distribuição espacial dos usos e coberturas urbanos, antropizados, florestais, hídricos e de solos expostos. Dessa forma, verificou-se que as mobilizações para a construção de rodovias, infraestrutura petrolífera e atividade agropecuária são fatores a serem considerados e regularizados para a proteção da cobertura florestal no PNY. O processamento digital de imagens no GEE, assim como a classificação por meio do algoritmo *Random Forest*, permitiu um mapeamento eficaz e eficiente, especialmente para facilitar o estudo de áreas com alta cobertura de nuvem. O código desenvolvido no GEE pode ser replicado, alterando-se apenas as variáveis de entrada como área de estudo, amostras e datas de análise.

5. Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001.

6. Referências

Bass, M. S., Finer, M., Jenkins, C. N., et al. (2010). Global conservation significance of Ecuador's Yasuní National Park. *PLoS ONE*, v. 5, n. 1.

- Breiman, L. (2001). Random Forests. *Machine Learning*, v. 45, p. 5–32.
- Finer, M., Vijay, V., Ponce, F., Jenkins, C. N. and Kahn, T. R. (jul 2009). Ecuador's Yasuní Biosphere Reserve: a brief modern history and conservation challenges. *Environmental Research Letters*, v. 4, n. 3, p. 034005.
- Gray, C. and Bilsborrow, R. (2020). Stability and change within indigenous land use in the Ecuadorian Amazon. *Global Environmental Change*, v. 63, n. August 2019, p. 102116.
- Kingsbury, D. V., Kramarz, T. and Jacques, K. (4 may 2019). Populism or Petrostate?: The Afterlives of Ecuador's Yasuní-ITT Initiative. *Society & Natural Resources*, v. 32, n. 5, p. 530–547.
- Llerena-Montoya, S., Velastegui-Montoya, A., Zhirzhan-Azanza, B., et al. (2021). Multitemporal Analysis of Land Use and Land Cover within an Oil Block in the Ecuadorian Amazon. *ISPRS International Journal of Geo-Information*, v. 10, n. 3, p. 191.
- Lovejoy, T. E. and Nobre, C. (20 dec 2019). Amazon tipping point: Last chance for action. *Science Advances*, v. 5, n. 12, p. eaba2949.
- Ministerio del Ambiente (2011). Plan de Manejo del Parque Nacional Yasuní.
- Olivares, I., Svenning, J.-C., Van Bodegom, P. M., Valencia, R. and Balslev, H. (mar 2017). Stability in a changing world - palm community dynamics in the hyperdiverse western Amazon over 17 years. *Global Change Biology*, v. 23, n. 3, p. 1232–1239.
- Parente, L., Taquary, E., Silva, A. P., Souza, C. and Ferreira, L. (2019). Next generation mapping: Combining deep learning, cloud computing, and big remote sensing data. *Remote Sensing*, v. 11, n. 23.
- Peel, M. C., Finlayson, B. L. and McMahon, T. A. (2007). Updated world map of the Köppen-Geiger climate classification. *Hydrology and earth system sciences*, v. 11, p. 1633–1644.
- Sellers, S., Bilsborrow, R., Salinas, V. and Mena, C. (2017). Population and development in the Amazon: A longitudinal study of migrant settlers in the Northern Ecuadorian Amazon. *Acta Amazonica*, v. 47, n. 4, p. 321–330.
- Tamiminia, H., Salehi, B., Mahdianpari, M., et al. (jun 2020). Google Earth Engine for geo-big data applications: A meta-analysis and systematic review. *ISPRS Journal of Photogrammetry and Remote Sensing*, v. 164, n. March, p. 152–170.
- Thieme, A., Hettler, B. and Finer, M. (2018). Deforestación Petrolera en el Parque Nacional Yasuní, Amazonía Ecuatoriana. MAAP. <https://www.arcgis.com/apps/Cascade/index.html?appid=2ee1d20d085543ed82a6fceb8cf91e20>.
- Thonfeld, F., Steinbach, S., Muro, J. and Kirimi, F. (25 mar 2020). Long-Term Land Use/Land Cover Change Assessment of the Kilombero Catchment in Tanzania Using Random Forest Classification and Robust Change Vector Analysis. *Remote Sensing*, v. 12, n. 7, p. 1057.
- UN Environment Programme World Conservation Monitoring Centre (2021). Explore the World's Protected Areas. <https://www.protectedplanet.net/en>.

Vazios de Informação & Informalidade: Web-Scraping para Identificação de Assentamentos Precários

Ana Carolina C. Carneiro¹, Flávia da Fonseca Feitosa¹, Luis Felipe B. da Cunha ¹

¹Laboratório de Estudos e Projetos Urbanos e Regionais (LEPUR) - Programa de Pós-Graduação em Planejamento e Gestão Territorial da Universidade Federal do ABC (UFABC) – São Bernardo do Campo, SP - Brasil

carolina.carneiro@ufabc.edu.br, flavia.feitosa@ufabc.edu.br, luis.cunha@ufabc.edu.br

Abstract. *This article presents an exploratory study to test the use of data from the construction of alternative web models for the identification of precarious settlements. The informal character of real estate transactions in AP, in which AP properties predominate, to the low density and absence of specialized property advertisements on the web, associated with the presence of AP. Thus, such “information gaps” represent information relevant to the identification of AP. The methodology used consists of mining 21,067 real estate advertisements in the city of São Bernardo do Campo (SP) on the ImóvelWeb portal. From the data, logistic models were built to estimate the probability of PA presence. The tests carried out confirmed that, as dynamic, the formal real estate markets were formed in a localized way in the territory and mostly in the perimeters of the AP, confirming an initial hypothesis*

Resumo. *Este artigo apresenta um estudo exploratório para testar a utilização de dados alternativos, provenientes de web-scraping na construção de modelos voltados para a identificação de assentamentos precários (AP). Dado o caráter informal das transações imobiliárias em AP, nos quais predominam situações de insegurança da posse, partiu-se da hipótese de que áreas caracterizadas por baixa densidade e até mesmo ausência de anúncios de imóveis em portais web especializados, estão associadas à presença de AP. Assim, tais “vazios de informação” representam, na verdade, informação relevante para a identificação de AP. A metodologia utilizada consistiu na mineração de 21.067 anúncios imobiliários da cidade de São Bernardo do Campo (SP) no portal ImóvelWeb. A partir desses dados, foram construídos modelos logísticos para estimar a probabilidade de presença de AP. Os testes realizados confirmaram que as dinâmicas do mercado imobiliários formal se dão de forma concentrada no território e majoritariamente fora dos perímetros dos AP, confirmando a hipótese inicial do trabalho.*

1.Introdução

Os assentamentos precários (AP) diferem ao longo de todo o território nacional no que se refere à tipologia, métodos construtivos, localização, cultura e nível de precariedade. Este fato dificulta sua identificação e caracterização com base em um único método para todo o país, o que impacta o desenvolvimento de políticas públicas voltadas para a regularização fundiária e programas habitacionais (DENALDI et al., 2020). Em que pese a relevância dos problemas habitacionais brasileiros, o país apresenta escassez de

informação sobre seus AP, o que vem mobilizando o desenvolvimento de distintas metodologias para sua identificação e caracterização – (MARQUES et al., 2007); (DENALDI; FEITOSA, 2016); (FEITOSA et al., 2019); (FEITOSA et al., no prelo). Em âmbito internacional, os AP também mobilizam esforços de identificação, com destaque para a pesquisa de (MAHABIR et al., 2020) de Detecção e Mapeamento de favelas utilizando dados abertos no Quênia, que buscou identificar fontes de dados abertos alternativos, incluindo a mineração de dados imobiliários, que poderiam ser associados a dados de sensoriamento remoto para identificar favelas com carência de dados.

O presente trabalho busca verificar se dados imobiliários extraídos de portais web podem ser úteis no contexto brasileiro, contribuindo para o desenvolvimento das metodologias existentes. Com o alcance cada vez maior da internet, aumenta-se o volume, variedade e velocidade de dados gerados a todo instante, os quais podem subsidiar inúmeras aplicações. Cabe salientar, entretanto, que estes dados – muitos deles denominados “big data” – apresentam limitações, entre as quais destaca-se o fato de estarem restritos a usuários e mercados (SHEARMUR, 2015). Ao tratar do uso destes dados em estudos urbanos, Feitosa (2020) ressalta que tal característica pode restringir nossa visão de “urbano”, assim como de sociedade. A autora afirma que estes dados são tendenciosos por natureza e ignoram pessoas que estão alheias aos mercados e atividades específicas que estão sendo rastreadas; excluem boa parte da sociedade e contribuem para deixar invisível aqueles que mais precisam de visibilidade (Feitosa, 2020, p. 4). Ao mesmo tempo, a vasta gama de dados geoespaciais disponíveis, criam um potencial para geovisualização capaz de apoiar o planejamento urbano voltado para esta parcela invisível da sociedade (MACEACHREN, 2004)

Buscando dar outro sentido para estas considerações, o presente trabalho, parte da hipótese de que a baixa concentração ou até mesmo a ausência de dados pode representar informação relevante para a identificação de áreas precárias. Para tanto, apresenta um estudo exploratório similar ao realizado por (MAHABIR et al., 2020) para a cidade de São Bernardo do Campo. O estudo testa a utilização de dados imobiliários extraídos de portais web de imóveis (*web scraping*¹) na construção de modelos para identificar AP. Acredita-se que, ao construir uma representação da ‘formalidade urbana’, constituída pela concentração de imóveis anunciados em portais web por imobiliárias, seja possível identificar vazios de informação que caracterizam a informalidade. Busca-se, assim, verificar se tais vazios representam um aspecto territorial relevante para estudos sobre a precariedade habitacional.

2.Método

O estudo consistiu na construção de modelos de regressão logística para a identificação de AP a partir da densidade de endereços de imóveis anunciados na internet. A regressão logística é um modelo estatístico formulado para prever e explicar uma variável categórica binária (dependente) a partir da sua relação com variáveis independentes.

¹ Web scraping é a coleta automatizada de dados da internet, também conhecida como raspagem de dados ou mineração de dados em português (MITCHELL, 2018).

As etapas do estudo foram: 1) coleta de dados; 2) tratamento e integração dos dados; 3) criação de superfícies de densidade Kernel; 4) construção de modelos de regressão logística; 5) Mapeamento da superfície de probabilidade².

A coleta de dados primários foi realizada através da mineração de dados de anúncios imobiliários do portal ImóvelWeb (“imovelweb”, 2022), entre os meses de julho e agosto de 2022, com o auxílio do software Power Automate. Os anúncios coletados foram tratados, filtrados e geocodificados com o auxílio da Interface de Programação de Aplicação (API) do Google Planilhas. Em seguida, foram estimadas duas superfícies de densidade kernel a partir destes dados, ambas com largura de banda de 150, sendo uma delas simples (Dens_Simples) e a outra ponderada pelo valor do imóvel anunciado (Dens_Pond). Estas superfícies de densidade de anúncios de imóveis serviram como variáveis independentes dos modelos logísticos.

Para representar a presença de AP (variável dependente dos modelos), considerou-se o mapeamento realizado pelo Diagnóstico Habitacional da Região do Grande ABC – DHABC (DENALDI; FEITOSA, 2016). Este levantamento identifica quatro tipologias de AP:

- T1: assentamentos urbanizados, consolidados e irregulares, que demandam ações de regularização fundiária;
- T2: AP, irregulares e consolidáveis, que demandam obras de infraestrutura, podendo necessitar de alguma remoção;
- T3: AP, irregulares e consolidáveis, que demandam obras complexas de urbanização e/ou percentual elevado de remoção;
- T4: AP, irregulares e não consolidáveis, cuja solução é a remoção total dos domicílios.

A partir dos kernels, foram construídos quatro modelos de regressão logística com o auxílio do Software R para estimar a probabilidade da presença de AP a partir da densidade de anúncios imobiliários, considerando-se a densidade de ocupação das células (Quadro 1).

Para construção dos modelos, considerou-se duas alternativas de variável dependente: a primeira inclui apenas as tipologias T2, T3, T4 (AP_234), ao passo que a outra inclui todas as tipologias, inclusive a T1, que abarca assentamentos já urbanizados, que aguardam apenas regularização (AP_1234). As variáveis geradas foram integradas em uma base celular vetorial com resolução de 100m e utilizadas para a construção de 4 modelos de regressão logística, conforma apresentado no Quadro 1.

Quadro 1 – Variáveis consideradas nos modelos de regressão logística

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Variável Dependente (Y)	AP_1234	AP_234	AP_1234	AP_234
Variável Independente (X)	Dens_Simples	Dens_Simples	Dens_Pond	Dens_Pond

² Superfícies de probabilidade são obtidas pela regressão logística e apresentam uma visão mais informativa do que apenas o mapa de classificação (sim ou não), obtido pela análise discriminante (FEITOSA et al., 2019).

Os modelos de regressão logística foram construídos para estimar a probabilidade de uma área (célula) ser (1) ou não (0) um AP a partir de sua relação a densidade de imóveis anunciados em portais web. Para cada modelo de regressão logística foram geradas e mapeadas superfícies de probabilidade.

3.Resultados

A mineração de dados obteve um retorno de 58.007 anúncios, dos quais 21.067 puderam ser geocodificados, por possuírem endereços completos. Os anúncios válidos e geocodificados são apresentados na Figura 1. A partir destes dados gerou-se a estimativa de densidade Kernel apresentada na Figura 2, na qual se encontram sobrepostos os assentamentos precários de São Bernardo do Campo (DENALDI; FEITOSA, 2016). Os resultados apresentados na Figura 3 apontam uma concentração dos anúncios imobiliários em áreas sem assentamentos precários.

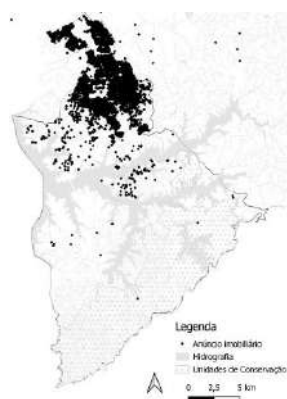


Figura 1- Mapeamento dos anúncios imobiliários coletados para São Bernardo do Campo

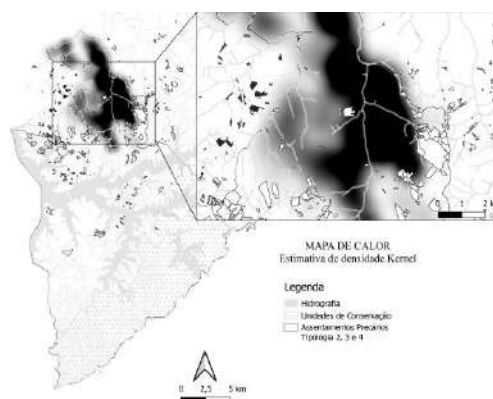


Figura 2 – Estimativa de densidade Kernel

Os resultados dos quatro modelos logísticos estimados são apresentados na Tabela .

Tabela 1 - Coeficientes Estimados

	Modelo 1	Modelo 2	Modelo 3	Modelo 4
Constante	-1,566 ***	-2,058 ***	-1,563 ***	-2,057 ***
Dens_Simples	-0,044 ***	-0,056 ***	-	-
Dens_Pond	-	-	-6,546.10 ⁻⁸ ***	-7,852.10 ⁻⁸ ***
Erro Padrão (EP)	0,016	0,019	1,605.10 ⁻²	1,922.10 ⁻²
AIC	26552	20176	26515	20158

*** indica nível de significância estatística de 0,001

Em todos os modelos, os coeficientes estimados para a variável independente (Dens_Simples ou Den_Ponderada) foram negativos e significativos, o que corrobora a

hipótese inicial de que os dados de anúncios de imóveis estão negativamente correlacionados com a presença de assentamentos precários.

O Akaike Information Criterion (AIC) fornece um método para avaliar a qualidade do modelo por meio de comparação com outros relacionados, sendo o mais indicado aquele que apresentar o menor coeficiente. No caso deste estudo, os modelos que apresentaram os melhores resultados foram os que desconsideraram a tipologia 1 de AP na análise (Modelos 2 e 4), restringindo-se aos assentamentos mais precários. Entre os dois modelos, o Modelo 4, que considera a superfície de densidade kernel ponderada pelos preços dos imóveis (Dens_Pond) apresentou resultados melhores.

A Figura 4 apresenta a superfície de probabilidade de presença de AP gerada a partir dos resultados do Modelo 4. No mapa da Figura 3, as células mais escuras estão mais próximas de 1 e apresentam maiores probabilidades de serem AP, ao passo que as células mais claras apresentam menores probabilidades. Uma análise qualitativa foi conduzida a partir da sobreposição dos polígonos dos AP (em vermelho), indicando a concentração de AP em áreas escuras, ou seja, apontadas pelo modelo como de baixa probabilidade.

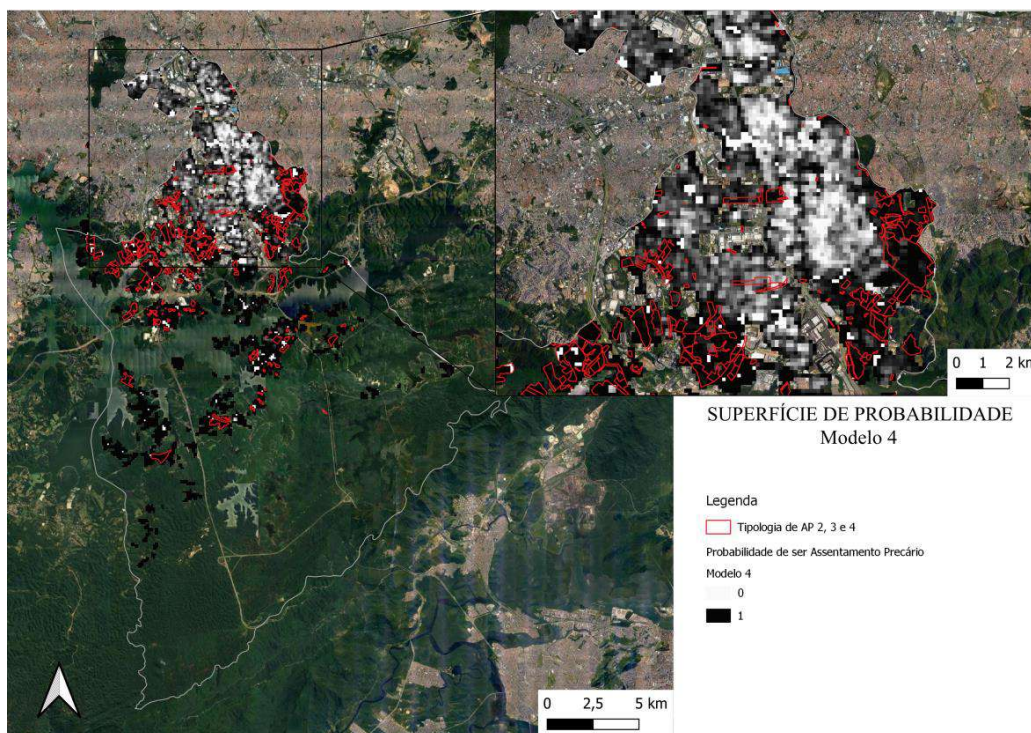


Figura 3: Mapa de superfície de probabilidade para o modelo 4

4. Considerações Finais

A identificação de assentamentos precários é uma tarefa essencial na elaboração de programas habitacionais, definição de prioridades e adequação de financiamento à distintas realidades. Apesar de seu carácter essencial, às metodologias de identificação de AP são fortemente impactadas pelas inúmeras limitações existentes na produção de informações e até mesmo da escassez de informações (FEITOSA et al., 2019). Dentro

deste cenário, este estudo se mostrou eficaz ao trazer a possibilidade de uma nova variável que possa auxiliar na identificação de assentamentos precários brasileiros. Devido ao aumento exponencial da produção de dados e o seu carácter desigual dentro do território, verifica-se que é possível extrair informações importantes através da espacialização destes dados, incluindo, os vazios informacionais gerados.

Ao analisar a relação entre a densidade de imóveis anunciados em portais web e a presença de assentamentos precários, verifica-se que a movimentação do mercado formal em ambiente online é concentrada no tecido urbano e não engloba grandes porções da periferia da cidade, onde, geralmente, encontram-se os assentamentos precários. Os resultados encontrados, corroboram a hipótese inicial.

Pelo carácter exploratório do estudo, entende-se que há uma vasta possibilidade de aperfeiçoamento da abordagem proposta, como o aumento do tempo de coleta de dados de modo a capturar as dinâmicas espaço-temporais do mercado imobiliário formal online e a elaboração de modelos utilizando outras variáveis, como renda e uso e ocupação do solo.

Referências

- DENALDI, R. et al. Produto 1 – Estudo conceitual e metodológico: Pesquisa de Núcleos Urbanos Informais no Brasil. [s.l.] SNH/MDR e IPEA, 6 abr. 2020.
- DENALDI, R.; FEITOSA, F. Diagnóstico Habitacional Regional do Grande ABC.pdf. Santo André: Consórcio Intermunicipal Grande ABC; Universidade Federal do ABC, set. 2016.
- FEITOSA, F. et al. MAPPA: Metodologia para Identificação e Caracterização de Assentamentos Precários em Regiões Metropolitanas Paulistas. ABC: UFABC, 2019.
- FEITOSA, F. et al. Modelagem para a Identificação de Núcleos Urbanos Informais: Uma Proposta Metodológica. Em: Núcleos Urbanos Informais - abordagens territoriais da irregularidade fundiária e da precariedade habitacional. Brasília: [s.n.].
- imovelweb. Portal Imobiliário. Disponível em: <<https://www.imovelweb.com.br/>>.
- MACEACHREN, A. et al. Geovisualization for Knowledge Construction and Decision Support. IEEE Computer Graphics and Applications, 2004
- MAHABIR, R. et al. Detecting and mapping slums using open data: a case study in Kenya. International Journal of Digital Earth, v. 13, n. 6, p. 683–707, 2 jun. 2020.
- MARQUES, E. et al. (EDS.). Assentamentos precários no Brasil urbano. São Paulo, Brazil]: [Brasília, Brazil: Centro de Estudos da Metrópole / CEBRAP; Secretaria Nacional de Habitação / Ministério das Cidades, 2007.
- MITCHELL, R. E. Web scraping with Python: collecting more data from the modern web. Second edition ed. Sebastopol, CA: O'Reilly Media, 2018.
- SHEARMUR, R. Dazzled by data: Big Data, the census and urban geography. Urban Geography, v. 36, n. 7, p. 965–968, 3 out. 2015.

Integrando Métodos de Análise Durante a Coleta de Amostras de Uso e Cobertura da Terra

Abner Ernâni dos Anjos¹, Karine Reis Ferreira¹, Gilberto Ribeiro de Queiroz¹,
Fabiana Zioti¹, Gabriel Sansigolo¹

¹ Instituto Nacional de Pesquisas Espaciais (INPE),
Av. dos Astronautas 1758 – 12227-010, São José dos Campos – SP, Brasil

{abner.anjos, karine.ferreira, gilberto.queiroz,
fabiana.zioti, gabriel.sansigolo}@inpe.br

Abstract. *The Land Use and Land Cover (LULC) mapping based on machine learning requires collecting samples for the training of predictive models. This step is a time-consuming and expensive process, in addition this activity is subject to mislabeling samples. Therefore, several methods for quality assessment have been developed and applied successfully, but attaching the collect process with assessment becomes a challenge on some platforms. This article presents a work in progress that aims to integrate methods for sample analysis during the collection process in the TerraCollect system under development in the Brazil Data Cube.*

Resumo. *A produção de mapas de uso e cobertura da terra com base em aprendizado de máquina requer a coleta de amostras para treinamento de modelos preditivos. Esta etapa é custosa e pode demandar tempo, além de ser uma atividade sujeita a erros na rotulação. Por conta disso, abordagens para análise da qualidade de dados de treinamento têm sido desenvolvidas e empregadas com sucesso, porém unir estas duas atividades ainda é um desafio em diversas plataformas. Este artigo apresenta um trabalho em andamento que busca a integração de métodos para a análise das amostras durante a coleta na plataforma web TerraCollect em desenvolvimento no projeto Brazil Data Cube.*

1. Introdução

Mapas de uso e cobertura da terra, em inglês *Land Use and Land Cover* (LULC), são importantes para representar a ação humana no meio ambiente. Esses mapas são indispensáveis para órgãos governamentais que procuram desenvolver políticas públicas ambientais e territoriais eficazes para prevenir altos impactos no meio-ambiente [Hansen and Loveland 2012]. Esses mapas são produzidos principalmente a partir de imagens de sensoriamento remoto. Atualmente, diferentes missões estão lançando satélites de observação da Terra e produzindo grandes volumes de dados abertos de imagens consistentes ao longo do tempo e do espaço.

A análise de séries temporais provenientes do empilhamento de imagens de satélite combinadas com métodos de aprendizado de máquina é uma técnica amplamente usada para produzir mapas de LULC [Simoes et al. 2020]. A maioria desses métodos são supervisionados, ou seja, necessitam de um conjunto representativo de amostras rotuladas a ser subdividido em treinamento, validação e teste [Santos et al. 2021, Goodfellow et al. 2016]. Portanto, a qualidade das amostras de treinamento é crucial no processo de classificação, pois leva a mapas com melhor acurácia.

Coletar amostras de treinamento é um processo custoso e pode demandar uma boa parcela do tempo do grupo de especialistas envolvidos. Em grandes áreas, a variabilidade das classes de LULC é alta e intrínseca em diferentes regiões e períodos devido à heterogeneidade da biodiversidade, bem como condições climáticas e práticas de manejo distintas. Por isso, as amostras, sejam coletadas em campo ou por interpretação visual a partir de imagens de alta resolução, podem conter ruídos e também estão sujeitas a erros na rotulação [Santos et al. 2021]. A coleta pode ser realizada através do uso de Sistemas de Informação Geográfica (SIG) de propósito geral, como o QGIS, ou por sistemas que foram desenvolvidos especificamente para este processo, como o *CollectEarth*.

O uso destes sistemas facilita a avaliação pois disponibilizam métodos de processamento de imagens e geometrias [Rwanga et al. 2017]. Enquanto SIG's de propósito geral possuem diversas ferramentas analíticas que exigem experiência e conhecimento específico para a configuração, os sistemas de coleta possuem funcionalidades mais restritas e objetivas e não integram muitas ferramentas analíticas.

Diversos métodos para melhorar a qualidade do conjunto de amostras têm sido desenvolvidos e empregados com sucesso [Santos et al. 2021, Tuia et al. 2009, Rwanga et al. 2017, Wickham and Golemund 2017]. No entanto, as ferramentas existentes para coleta não integram esses métodos para auxiliar na produção de amostras representativas [Santos et al. 2021].

Este trabalho apresenta uma aplicação que integra diferentes abordagens para análise de amostras de LULC que auxiliam os especialistas durante e após a atividade de coleta. Essa aplicação está sendo implementada como uma extensão da plataforma *TerraCollect* em desenvolvimento no projeto *Brazil Data Cube* [Ferreira et al. 2020]. As abordagens incluem a Análise Exploratória de Dados [Wickham and Golemund 2017], Estimativa de Probabilidades, *Active Learning* [Tuia et al. 2009] e *Class Noise Reduction* [Santos et al. 2021].

2. Arquitetura Geral

TerraCollect é uma plataforma *web* para coleta de amostras de LULC que está sendo desenvolvida no projeto Brazil Data Cube (BDC). Esta plataforma utiliza os serviços *web* do BDC para descoberta e acesso a dados de observação da Terra, como por exemplo o *Spatiotemporal Asset Catalog* (STAC), *Web Time Series Service* (WTSS) e *Web Land Trajectory Service* (WLTS).

O *TerraCollect* fornece a visualização de imagens de sensoriamento de cubos e coleções de dados produzidas pelo BDC a partir dos satélites *CBERS-4*, *CBERS-4A*, *Landsat-8*, *Sentinel-2* e *MODIS* catalogados com o STAC. A plataforma também permite o acesso às séries temporais e trajetórias de LULC, respectivamente, com os serviços WTSS e WLTS para auxiliar os especialistas na tomada de decisão durante a coleta.

Após a coleta, as amostras são salvas em um banco de dados utilizando o modelo do *Sample Database Model* (Sample-DB) através do serviço *Sample Web Service* (Sample-WS). As amostras são representadas por seus atributos temporais (*start_date* e *end_date*), sua localização espacial (longitude e latitude) e a sua classe LULC. Com esses atributos, a extensão para a análise de amostras utiliza o SITS para a extração de séries temporais e para execução dos métodos de análise. A Figura 1 apresenta a arquitetura de *software* com a visão geral da integração desses métodos no *TerraCollect*.

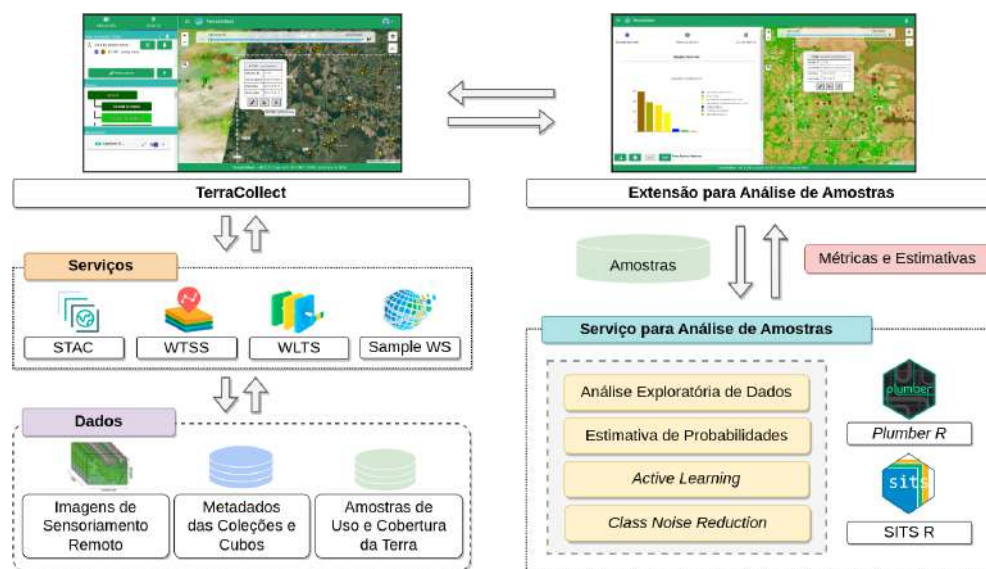


Figura 1. Arquitetura do TerraCollect com a extensão para a análise de amostras.

O pacote SITS fornece um conjunto de ferramentas para a extração, análise e classificação de séries temporais com aprendizado de máquina. O SITS adapta algoritmos como o *Random Forest*, *Support Vector Machines* (SVM), *Convolutional Neural Networks* (CNN's), dentre outros para trabalhar com este formato de dados [Simões 2021]. Com estas ferramentas é possível implementar as abordagens para a análise de amostras estendendo a aplicação do TerraCollect. A seguir são apresentados os dois principais componentes da extensão de análise de amostras.

2.1. Extensão para a Plataforma TerraCollect

Com as amostras disponíveis, os usuários podem acessar a extensão pela aplicação *web* do TerraCollect. Esta aplicação foi desenvolvida com o *framework Angular* [Jain et al. 2014] para criar uma interface gráfica com visualizações interativas de resultados. A extensão possui componentes para a execução das abordagens implementadas no SITS: Análise Exploratória de Dados, Estimativa de Probabilidades, *Active Learning* e *Class Noise Reduction*. Para integrar essas abordagens em linguagem de programação R na aplicação TerraCollect foi desenvolvido um serviço especializado para análise de amostras.

2.2. Serviço para Análise de Amostras

O serviço para análise de amostras é uma *Application Programming Interface* (API) desenvolvida com o pacote chamado *Plumber R* [Schloerke and Allen 2022]. Uma API permite a comunicação entre diferentes aplicações e serviços *web* [Fowler and Lewis 2014]. Para este trabalho um dos requisitos essenciais é a comunicação entre a aplicação TerraCollect e o SITS.

O pacote *Plumber R* converte as funções implementadas em *scripts* R pré-existente em uma API usando uma coleção de comentários especiais para encapsulá-los nos métodos *GET*, *POST* e *DELETE* [Schloerke and Allen 2022]. Desta forma o TerraCollect envia as requisições para executar a extração de séries temporais e a análise de amostras no SITS.

3. Abordagens para a Análise de Amostras

Essa seção apresenta a descrição das abordagens para a análise de amostras que serão integradas na arquitetura proposta.

3.1. Análise Exploratória de Amostras

A Análise Exploratória de Dados (AED) é um método usado para resumir as principais características dos dados, geralmente usando técnicas de visualização com gráficos ou tabelas. Desta forma é possível identificar padrões, detectar anomalias e verificar se os dados atendem aos objetivos esperados ou se necessitam de ajustes [Wickham and Grommund 2017]. Na extensão de análise de amostras, o componente para a AED fornece o gráfico da distribuição de frequência de amostras por classe e a visualização do padrão das séries temporais com as bandas e índices de vegetação para resumir o conjunto de amostras disponíveis. O SITS usa uma *Generalized Additive Models* (GAM) para suavizar a série temporal, e assim, estimar um padrão de comportamento no período. Com essas visualizações o analista consegue distinguir as características de cada classe.

3.2. Estimativa de Probabilidades

As amostras armazenadas no banco de dados do *TerraCollect* podem ser usadas como um conjunto de dados base de treinamento para um modelos de aprendizado de máquina baseados em *RandomForest*, SVM's ou CNN's de forma supervisionada. Desta forma, durante o processo de coleta de novos dados, esses modelos podem ser utilizados para inferir a probabilidade de uma nova amostra pertencer a uma dada classe [E. Walpole et al. 2012]. No entanto, o uso do resultado da predição pode gerar problemas de *overfitting* resultando na especificação do modelo. Por essa razão, estudos usando a incerteza dos resultados são necessários como as técnicas de *Active Learning* apresentadas por Tuia et al. 2009.

3.3. Active Learning

Active Learning usa um modelo de aprendizado de máquina pré-treinado com um subconjunto do treinamento para analisar as divergências nas previsões do modelo ao prever novas amostras [Tuia et al. 2009]. Este método é derivado do mesmo princípio da estimativa de probabilidades, porém é realizado um cálculo com base na predição como a entropia e a margem de confiança que medem a incerteza para avaliar a representatividade de uma amostra.

3.4. Class Noise Reduction Method

Santos et al. 2021 propôs um método que avalia a qualidade das amostras de LULC com base na inferência *bayesiana* aplicada no agrupamento com o método *Self-organizing maps* (SOM). Como apresentado no estudo, os vizinhos de cada neurônio em um mapa SOM fornecem informações sobre a variabilidade intraclasse e interclasse, desta forma a inferência é aplicada em cada um dos neurônios para mensurar a probabilidade do rótulo ter sido atribuído corretamente. Assim, é feita a análise com base nos filtros de *threshold* para a probabilidade *a priori* e *posteriori* resultando na classificação das amostras em *status* de “*clean*” (permanece no conjunto), “*analysis*” (análise necessária) e “*remove*” (remoção aconselhável). O componente para a execução desse método possui a visualização do mapa SOM onde o analista consegue explorar os neurônios e recuperar as amostras de cada grupo. Com a amostra selecionada pode-se analisar a série temporal e compará-la com o padrão da classe.

4. Resultados Parciais

Foi realizado um estudo de caso no Bioma Cerrado com 591 amostras de agricultura e floresta extraídas dos mapas do *TerraClass*. Essas amostras foram fornecidas pela Embrapa para re-classificação com base na interpretação visual das imagens de satélite com o auxílio da plataforma *TerraCollect* para o período de Agosto de 2019 à Agosto de 2020. Para análise foram usadas as séries temporais do cubo de imagens do *CBERS-4* sensor *WFI* com uma resolução espacial de 64 metros e 16 dias de resolução temporal usando os índices de vegetação *Enhanced Vegetation Index* (EVI) e *Normalized Difference Vegetation Index* (NDVI).

A Figura 2.a mostra a tela inicial da ferramenta com as opções disponíveis para realizar a análise exploratória. Esta tela apresenta o gráfico de distribuição de frequência das amostras por classe destacado na Figura 2.b. A Figura 2.c apresenta em destaque a seleção de um padrão de séries temporais para uma dada classe de Agricultura.

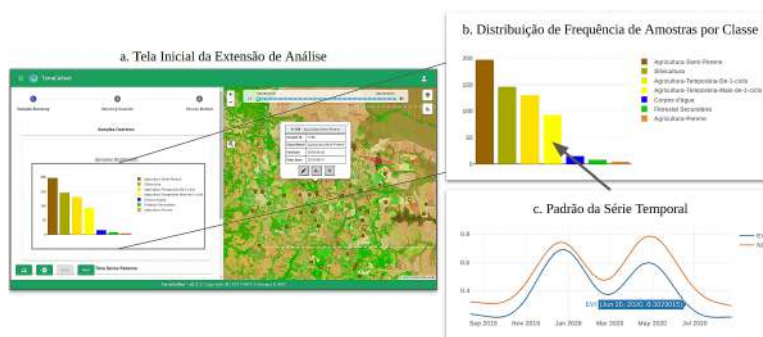


Figura 2. Tela inicial da extensão com a Análise Exploratória de Amostras.

Para fins de teste, foram selecionadas somente as amostras de agricultura das 591 amostras para treinar um modelo baseado no *Random Forest* com 200 árvores. Com o modelo treinado, foi selecionado uma nova amostra para estimar a probabilidade. A Figura 3.a apresenta o gráfico de pizza da predição resultando em 50% de chance de Agricultura semi-perene.

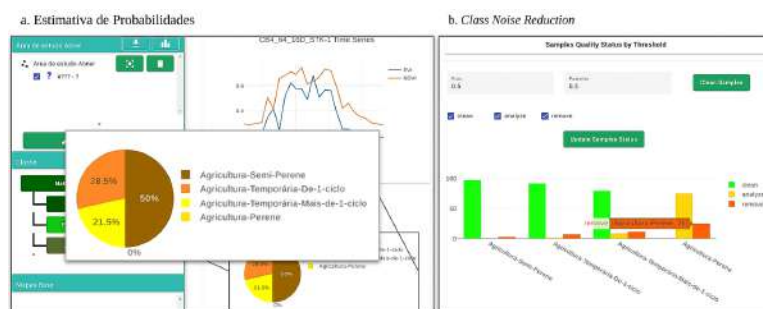


Figura 3. Recorte da tela da extensão com os resultados da análise.

Já a Figura 3.b apresenta o resultado do método *Class Noise Reduction* com gráfico da distribuição da quantidade de amostras por *status*. Para o estudo de caso, a classe de Agricultura Perene apresentou maior número de amostras ruidosas para análise e remoção. O analista pode selecionar as amostras e usar o *status* em conjunto com as ferramentas do *TerraCollect* (como o WLTS e WTSS) para tomar a decisão sobre quais dados devem permanecer no conjunto.

5. Considerações finais

Este artigo apresenta uma aplicação para integrar métodos de análise de amostras durante a atividade de coleta. A aplicação foi implementada como uma extensão da plataforma *TerraCollect* para a execução das abordagens implementadas no SITS. Os experimentos permitiram um estudo mais elaborado dos métodos de análise de amostras e a organização e implementação desses métodos em uma aplicação *web*. Porém, ainda é necessário realizar estudos e experimentos com as técnicas de *Active Learning* para demonstrar a sua viabilidade no refinamento de amostras.

Para trabalhos futuros, espera-se revisar as ferramentas para a extração e o armazenamento das séries temporais. Devido ao volume de dados usados nos experimentos, conclui-se que ainda há um alto custo de tempo para a extração. Dessa forma é necessário propor um novo modelo de banco de dados para o armazenamento e espera-se que futuramente a extração seja otimizada usando programação paralela ou demais técnicas relacionadas.

Referências

- E. Walpole, R., Myers, R. H., Myers, S. L., and Ye, K. (2012). *Probability & Statistics for Engineers & Scientists*. Pearson, 9 edition.
- Ferreira, K. R., Queiroz, G. R., Vinhas, L., and et.al. (2020). Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24).
- Fowler, M. and Lewis, J. (2014). *Microservices*.
- Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.
- Hansen, M. C. and Loveland, T. R. (2012). A review of large area monitoring of land cover change using Landsat data. *Remote Sensing of Environment*, 122:66–74.
- Jain, N., Bhansali, A., and Mehta, D. (2014). Angularjs: A modern mvc framework in javascript. *Journal of Global Research in Computer Science*, 5(12):17–23.
- Rwanga, S. et al. (2017). Accuracy Assessment of Land Use/Land Cover classification Using Remote Sensing and GIS. *International Journal of Geosciences*, 8(4):611–622.
- Santos, L. A. et al. (2021). Quality control and class noise reduction of satellite image time series. *ISPRS Journal of Photogrammetry and Remote Sensing*, 177:75–88.
- Schloerke, B. and Allen, J. (2022). *plumber: An API Generator for R*. <https://www.rplumber.io>, <https://github.com/rstudio/plumber>.
- Simoës, R. et al. (2020). Land use and cover maps for Mato Grosso State in Brazil from 2001 to 2017. *Scientific Data*, 34(7).
- Simões, R. (2021). *Land use and land cover classification of satellite image time series using machine learning*. PhD thesis, Instituto Nacional de Pesquisas Espaciais - INPE, São José dos Campos - SP - Brasil.
- Tuia, D. et al. (2009). Active Learning Methods for Remote Sensing Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 47(7):2218–2232.
- Wickham, H. and Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media, 1 edition.

Mapeamento colaborativo do nível de ruído em Guayaquil durante o COVID-19

Andrés Velastegui-Montoya^{1 2}, Juliana O. Marquez³, Geancarlo Guerrero-Cabrera¹, Yadira Jaramillo-Lindao¹, Malcon Mora-Araus¹

¹Facultad de Ingeniería en Ciencias de la Tierra (FICT)
ESPOL Polytechnic University - Guayaquil, Guayas - Ecuador

²Centro de Investigación y Proyectos Aplicados a las Ciencias de la Tierra (CIPAT)
ESPOL Polytechnic University - Guayaquil, Guayas - Ecuador

³Instituto de Ciência e Tecnologia (ICT),
Universidade Federal de São Paulo (UNIFESP), São José dos Campos, SP - Brazil

{dvelaste, geagcabr, yalijara, magamora}@espol.edu.ec

juliana.marquez@unifesp.br

Abstract. *The SARS-CoV-2 virus triggered a never-before-seen scenario worldwide. Ecuador declared a state of emergency to contain coronavirus transmission by applying social isolation policies and mobility limitations. During this period, noise pollution decreased in Guayaquil city. This study proposes the use of smartphones to collect noise data and generate soundscape maps of Guayaquil during different time slots. The results revealed traffic as the predominant type of noise, followed by vehicles, buses, and people's voices. This article demonstrates the feasibility of crowdsourcing for the collection of environmental variables.*

Resumo. *O vírus SARS-CoV-2 desencadeou um cenário nunca antes visto no mundo. No Equador, um Estado de Exceção foi declarado para tentar conter a transmissão do coronavírus, aplicando políticas de isolamento social e limitações de mobilidade. Nesse contexto, evidenciou-se uma diminuição da poluição sonora na cidade de Guayaquil, Guayas. Neste estudo, propõe-se o uso de smartphones para a coleta colaborativa de dados de ruídos e geração de mapas da paisagem sonora em Guayaquil, para diferentes faixas horárias. Os resultados revelaram que o tipo de ruído predominante é o tráfego de automóveis, seguidos pelas vozes das pessoas. Este artigo demonstra a viabilidade do crowdsourcing para a coleta de variáveis ambientais.*

1. Introdução

Em dezembro de 2019, a Organização Mundial da Saúde (OMS) foi alertada sobre novos casos de uma “pneumonia viral”, relatados pela Comissão de Saúde Municipal de Wuhan (China). Em 30 de janeiro de 2020, a OMS declarou o surto representado pelo novo coronavírus como uma Emergência de Saúde Pública de Importância Internacional (ESPII). Em 11 de fevereiro do mesmo ano, o Comitê Internacional de Taxonomia de Vírus (ICTV) nomeou o novo vírus como “coronavírus da síndrome respiratória aguda grave 2 (SARS-CoV-2)” e a OMS anunciou como “COVID-19” [World Health Organization 2020b]. Devido a rápida disseminação e mortalidade do vírus, em 11 de março de 2020, a OMS

anunciou um estado de emergência global, caracterizando uma pandemia e, como consequência, criando novos e maiores desafios aos sistemas de saúde pública e à economia dos países [Akin and Gözel 2020, World Health Organization 2020a].

Eventualmente, a chegada do novo coronavírus nos países da América Latina exigiu políticas de saúde pública para mitigar sua rápida transmissão entre a população. Com a finalidade de conter a transmissão do coronavírus, em 16 de março de 2020, foi declarado Estado de Exceção em todo o território equatoriano e políticas de isolamento social e limitações de mobilidade foram aplicadas [Secretaría General de Comunicación de la Presidencia 2020]. Apesar das restrições, o país testemunhou um número desconcertante de infecções pelo COVID-19. Em abril de 2020, o Equador ficou em primeiro lugar em mortes per capita na América do Sul, concentradas principalmente na cidade de Guayaquil [Ching-Ávalos et al. 2021, el Comercio 2020].

No mesmo mês, foi implantado um sistema de semaforização para regular a mobilização no país. Este sistema é baseado em codificação por cor, em que: a cor verde, representou livre acesso; amarelo, restrição de acesso em alguns casos; e vermelho, o cumprimento obrigatório de vários regulamentos, como toque de recolher entre 14h00 e 5h00 da manhã, suspensão do trabalho e classes presenciais, entre outras medidas [Agencia EFE 2020]. Posteriormente, a mobilidade tornou-se progressivamente mais flexível, o que permitiu a reabertura das atividades laborais, especialmente aquelas que não podem ser realizadas de maneira remota. Assim, a cidade de Guayaquil permaneceu na cor vermelha até 19 de maio de 2020 [Ministerio del Trabajo 2020]. A reativação das atividades e mobilidade na cidade, produziu um aumento progressivo dos níveis de ruído na paisagem sonora. Este ruído pode ser descrito como um som indesejado que afeta negativamente a vida humana ou animal e que é definido como “qualquer som indesejado gerado pelas atividades humanas, como o ruído emitido por meios de transporte, rodoviário, aéreo, atividade industrial, entre outros [Murphy and King 2022a, Murphy and King 2022b]. O ruído é um indicador socioeconômico espacial, que permite identificar o estado da atividade dos cidadãos [Ko et al. 2011]; Ovedepo et al. (2019). Em relação ao nível do ruído, a Agência de Proteção Ambiental recomenda manter o ruído ambiental abaixo de 70 decibéis (dB), para evitar perda auditiva induzida [Environmental Protection Agency 2016].

No Equador, as medidas de restrição aplicadas em 2020 em consequência da pandemia, causaram a diminuição dos níveis de ruído. Após a implementação do Estado de Exceção, a cidade de Guayaquil sofreu uma queda de 60% no ruído sísmico [Instituto Geofísico 2020]. Sob esse contexto, as questões de pesquisa deste projeto avaliaram o nível de ruído em Guayaquil sob condições parciais de restrição de mobilidade e a destruição espacial desses ruídos em diferentes setores urbanos da cidade.

Assim, em um cenário único de restrições para conter a transmissão do novo coronavírus, identificamos e compreendemos a composição dos mais importantes agentes produtores de ruído na Cidade de Guayaquil durante o isolamento social. Para tanto, desenvolvemos um mapa de ruído usando dados de poluição sonora coletados colaborativamente em smartphones (crowdsourcing), que auxiliara no trabalho das autoridades locais, pois permite avaliar as áreas e faixas horárias com a maior intensidade de ruído.

2. Metodologia

A área de estudo corresponde à cidade de Guayaquil ($1^{\circ}57' - 2^{\circ}17' \text{ S}$ e $79^{\circ}44' - 80^{\circ}05' \text{ O}$), localizada no litoral do Equador, capital da província de Guayas (Figura 1). Sendo esta uma área metropolitana com extensão aproximada de 545,51 km², da pela cidade de Guayaquil e parte das paróquias de Durán, Samborondón e Daule.

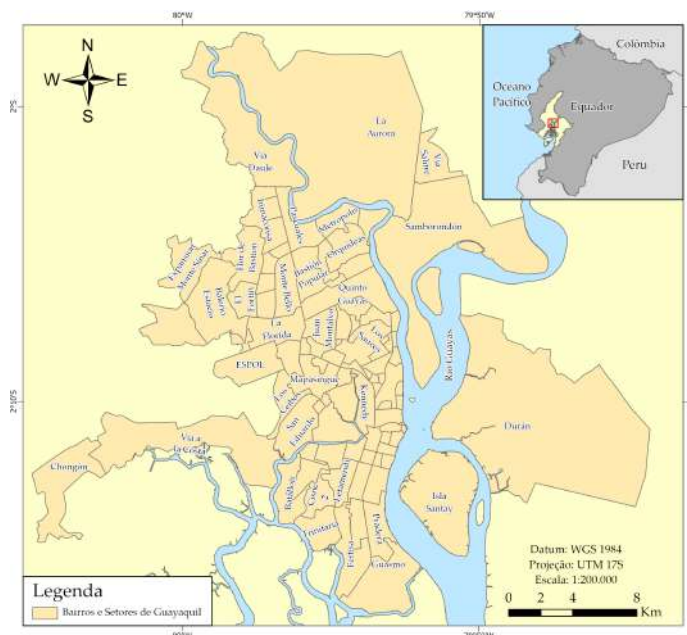


Figura 1. Área de estudo: cidade de Guayaquil (Guayas) e parte das paróquias de Durán, Samborondón e Daule.

Os cenários sonoros da cidade de Guayaquil foram caracterizados em quatro faixas horárias: 9h00, 13h00, 16h00 e 20h00. Do mesmo modo, os níveis de ruído máximo, mínimo e médio, foram registrados de forma colaborativa usando smartphones em diferentes regiões da cidade. Para o crowdsourcing de dados de ruídos com smartphones é necessário ter uma equipe voluntária disposta a medir e registrar o ruído ambiental e uma interface que permita o registro dos dados em tempo real.

A equipe voluntária deste projeto era composta por 70 estudantes da Escola Superior Politécnica do Litoral (ESPOL), que realizaram a atividade de medição como parte das horas necessárias em um dos estágios exigidos para a graduação. Dessa foram, os estudantes realizaram as seguintes atividades:

- Medir do ruído captado com a janela da residência aberta, por um período de 2 minutos, em um aplicativo de smartphone;
- Certificar-se de que nenhum ruído seja produzido dentro da casa durante os 2 minutos de medição, para garantir que a medição seja efetivamente o ruído externo;
- Registrar na interface digital os níveis coletados pelo aplicativo.

Devido a indisponibilidade de sonômetros para os 70 estudantes e, especialmente, devido ao prolongamento das medidas de restrição, foram utilizados aplicativos de smartphone para a medição do ruído, recurso este, já aplicado em outros estu-

dos [Murphy and King 2016]. Para os smartphones com sistema operacional iOS, utilizamos o App Decibel X: dB Sound Level Meter [SkyPaw 2022] e, para os com Android, o App Sonómetro: Sound Meter [Smart Tools 2022]. Durante as aplicações foram medidos os valores mínimo, médio e máximo do ruído em dB. A coleta dos dados foi realizada de forma contínua entre os dias 26 de maio e 2 de outubro de 2020, em quatro horários: 9h00, 13h00, 16h00 e 20h00.

A interface que possibilitou o registro dos dados de ruído ambiental em tempo real apresenta um formulário on-line criado na aplicação Survey 123 do software ArcGIS [Esri 2022b]. Este formulário permitiu armazenar a localização dos registros georreferenciados do ruído. Para isso, foi necessário inserir os níveis de mínimo, máximo e médio, além de especificar o tipo de ruído predominante durante o tempo de medição. Para melhorar a experiência de visualização e interação do usuário com os dados da paisagem sonora, desenvolvemos um visualizador de mapa em tempo real (<https://arcg.is/uaOrO>), que está vinculado com o formulário on-line de registro de dados. Este visualizador de mapa dinâmico foi criado no aplicativo Dashboard do ArcGIS [Esri 2022a]. Dessa forma, usuários especialistas ou não em SIG (Sistema de Informações Geográficas) podem visualizar e filtrar os dados da paisagem sonora da cidade, além de ter acesso aos mapas de níveis globais de ruído em diferentes faixas horárias.

3. Resultados e Discussões

O Dashboard (<https://arcg.is/uaOrO>) apresenta a paisagem sonora de Guayaquil no período de isolamento social causado pela pandemia da COVID-19, em 2020. Neste período, coletamos um total de 15.843 dados de níveis de ruído em dB, em diferentes pontos da cidade. Os dados coletados permitiram identificar as áreas de maior produção sonora e caracterizá-las segundo o tipo de ruído predominante. Ao comparar os dados dos primeiros e últimos 30 dias de medição, foi possível identificar que a cidade experimentou um acréscimo na poluição sonora, com variações nos valores médios de 2,1 dB (9h00), 1,5 dB (13h00), 1,8 dB (16h00) e 2,1 dB (20h00). Por outro lado, os valores máximos de ruído se mantiveram semelhantes, contudo, com valores superiores ao sugerido pela Agência de Proteção Ambiental. O trânsito de carros e ônibus (sem considerar motocicletas) e as vozes/gritos de pessoas na rua, foram estabelecidos como os produtores de ruído mais frequentes. A contribuição à poluição sonora da primeira fonte foi a predominante, com aproximadamente 12% dos registros, sofrendo uma diminuição durante as faixas da 16h00 e 20h00. Por outro lado, as vozes/gritos das pessoas na rua tiveram uma recorrência do 4% dos dados e prevaleceram nas faixas horárias diurnas e noturnas.

Os setores de Guayaquil que mais frequentemente apresentam altos níveis de poluição sonora são: San Felipe, Pradera, Orquídeas, Mucho Lote 2, Mapasingue, Saucés, Letamendi, Prosperina, Kennedy, Guayacanes, Garay, Fertisa, Febres Cordero, El Batallón, El Fortín, Durán, La Garzota, Carbo Concepción, Ceibos, Brisas del Río, Bellavista, Bastión Popular, Ayacucho, Aurora - Las Lojas y de Octubre Oeste. Ao considerarmos que as linhas de ônibus que percorrem Guayaquil possuem rotas convergentes em áreas como Orquídeas, 9 de Octubre, El Fortín e Mapasingue, entre outros bairros (Figura 1), associamos a este fator o aumento persistente dos níveis de ruído na cidade.

A coleta de dados realizada permitiu conhecer os valores do nível de ruído em

Guayaquil, em um cenário único de restrição parcial de mobilidade, que pode ser considerado um cenário ideal da cidade, sem maior intervenção de agentes produtores de ruído externo. Contudo, com a diminuição das restrições, a cidade Guayaquil retomou parcialmente suas atividades presenciais resultando em um acréscimo na poluição sonora da cidade.

4. Considerações Finais

Neste projeto, constatamos a viabilidade na utilização de dados de crowdsourcing para coleta e registro de variáveis ambientais com cidadãos treinados. Um grupo de 70 estudantes de graduação, mediu e registrou o ruído ambiental por meio de smartphones, em uma área urbana no Equador. Para garantir uma colaboração aberta dos usuários, foi essencial a utilização da aplicação Survey 123 do software ArcGIS para registro dos dados medidos nos aplicativos Apps Decibel X e Sonómetro. Além dos aplicativos utilizados para coleta de dados, outro componente fundamental da pesquisa é o visualizador de mapa dinâmico (Dashboard), que apresenta em tempo real os dados da paisagem sonora de Guayaquil.

Este projeto foi replicado no ano 2021, durante os mesmos meses considerados neste estudo, com a finalidade de comparar os dados coletados durante o isolamento social devido ao COVID-19 (ano 2020), com dados de um cenário quase normal, com menor isolamento social (ano 2021). Os resultados da etapa complementar do projeto será apresentada em outro trabalho, mas pelo momento está disponível um Story Map do projeto <https://storymaps.arcgis.com/stories/661c60d66b5442e48c47b9e133474ae8>. Diante dos resultados obtidos, entendemos que mediante adaptações na implementação de sonômetros automatizados para coleta de dados, é possível que este projeto possa ser aplicado em benefício das prefeituras na criação de mapas de ruídos da cidade.

Referências

- Agencia EFE (2020). Ecuador inicia en rojo su semáforo de flexibilización frente a la covid-19. <https://www.efe.com/efe/america/sociedad/ecuador-inicia-en-rojo-su-semaforo-de-flexibilizacion-frente-a-la-covid-19/20000013-4220315>. Acessado em 30 de agosto de 2022.
- Akin, L. and Gözel, M. G. (2020). Understanding dynamics of pandemics. *Turkish journal of medical sciences*, 50(9):515–519.
- Ching-Ávalos, S., Jaramillo-Lindao, Y., Velastegui-Montoya, A., Encalada Abarca, L., Chang-Silva, R., and Mosquera-Romero, M. (2021). Crowdsourcing of covid-19 symptoms map in ecuadorians. In *Proceedings of the LACCEI international Multi-conference for Engineering, Education and Technology*, pages Code–175610. Latin American and Caribbean Consortium of Engineering Institutions.
- el Comercio (2020). Guayas cierra abril del 2020 con más muertes que en 4 meses del 2019; el contexto del covid-19 en ecuador. <https://www.elcomercio.com/actualidad/record-muertes-guayas-coronavirus-ecuador.html>. Acessado em 30 de agosto de 2022.
- Environmental Protection Agency (2016). Epa identifies noise levels affecting health and welfare. <https://archive.epa.gov/epa/aboutepa/epa-identifies-noise-levels-affecting-health-and-welfare.html>. Acessado em 30 de agosto de 2022.

- Esri (2022a). Arcgis dashboards. environmental systems research institute. <https://www.esri.co/es-ec/productos/arcgis-dashboards/inicio>. Acessado em 30 de agosto de 2022.
- Esri (2022b). Arcgis survey123. environmental systems research institute. <https://survey123.arcgis.com/>. Acessado em 30 de agosto de 2022.
- Instituto Geofísico (2020). ¿ruido sísmico en ecuador se reduce debido a la restricción de movilidad! <https://www.igepn.edu.ec/interactuamos-con-usted/1809-ruido-sismico-en-ecuador-se-reduce-debido-a-la-restriccion-de-movilidad>. Acessado em 30 de agosto de 2022.
- Ko, J. H., Chang, S. I., and Lee, B. C. (2011). Noise impact assessment by utilizing noise map and gis: A case study in the city of chungju, republic of korea. *Applied Acoustics*, 72(8):544–550. Noise Mapping.
- Ministerio del Trabajo (2020). En guayaquil se retoma la atención con todas las medidas de seguridad. <https://www.trabajo.gob.ec/en-guayaquil-se-retoma-la-atencion-con-todas-las-medidas-de-seguridad/>. Acessado em 30 de agosto de 2022.
- Murphy, E. and King, E. A. (2016). Testing the accuracy of smartphones and sound level meter applications for measuring environmental noise. *Applied Acoustics*, 106:16–22.
- Murphy, E. and King, E. A. (2022a). Chapter 1 - understanding environmental noise pollution. In Murphy, E. and King, E. A., editors, *Environmental Noise Pollution (Second Edition)*, pages 1–7. Elsevier, Boston, second edition edition.
- Murphy, E. and King, E. A. (2022b). Chapter 2 - principles of environmental noise. In Murphy, E. and King, E. A., editors, *Environmental Noise Pollution (Second Edition)*, pages 9–51. Elsevier, Boston, second edition edition.
- Secretaría General de Comunicación de la Presidencia (2020). El presidente lenín moreno decreta estado de excepción para evitar la propagación del covid-19. [https://www.comunicacion.gob.ec/el-presidente-lenin-moreno-decreta-estado-de-excepcion-para-evitar-la-propagacion-del-covid-19/:.text=%E2%80%93%20La%20noche%20de%20este%20lunes,coronavirus%20\(COVID%2D19\)](https://www.comunicacion.gob.ec/el-presidente-lenin-moreno-decreta-estado-de-excepcion-para-evitar-la-propagacion-del-covid-19/:.text=%E2%80%93%20La%20noche%20de%20este%20lunes,coronavirus%20(COVID%2D19)). Acessado em 30 de agosto de 2022.
- SkyPaw (2022). Decibel x:db sound level meter. <https://apps.apple.com/us/app/decibel-x-db-sound-level-meter/id448155923>. Acessado em 30 de agosto de 2022.
- Smart Tools (2022). Sonómetro : Sound meter. <https://abrir.link/EHVLp>. Acessado em 30 de agosto de 2022.
- World Health Organization (2020a). Archived: Who timeline - covid-19. <https://www.who.int/news-room/detail/27-04-2020-who-timeline—covid-19>. Acessado em 30 de agosto de 2022.
- World Health Organization (2020b). Naming the coronavirus disease (covid-19) and the virus that causes it. [https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-\(covid-2019\)-and-the-virus-that-causes-it](https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/naming-the-coronavirus-disease-(covid-2019)-and-the-virus-that-causes-it). Acessado em 30 de agosto de 2022.

Previsão da Probabilidade de Fogo: ocorrências pós-alerta na Amazônia Maranhense

Ana Larissa Ribeiro de Freitas¹, Liana Oighenstein Anderson¹, Paulo Henrique Alves Leão¹, João Bosco Coura dos Reis¹, Celso Henrique L. Silva-Junior², Nathália S. Carvalho³, Ana Carolina M. Pessoa¹, Wanderson Henrique dos Santos¹

¹Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN), São José dos Campos - SP.

²University of California, Los Angeles - CA.

³Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos - SP.

{alarisig}@gmail.com, liana.anderson@cemaden.gov.br,
henri.leaos@gmail.com, joaodosreis89@gmail.com, celsohlsj@gmail.com,
nathalia.bioufla@gmail.com, acmoreirapessoa@gmail.com,
wan.mecanica@hotmail.com

Abstract. Fires and burning are threats to the lives of populations, ecosystems and the climate. Initiatives such as the Risk Management and Impacts of Forest Fires (CEMADEN) platform can contribute to reducing these risks. This research aimed to identify the behavior of the occurrence of hotspots compared to the predicted in the CEMADEN product. It was identified that the highest occurrence of hotspots, with 83.5%, is concentrated in the first four quarters of forecast (ASO, SON, OND and NDJ), and the decrease in alert levels from the DJF forecast. This analysis also contributes to the understanding of the alert dynamics and helps in prioritizing municipalities for fire prevention actions.

Resumo. Os incêndios e queimadas são ameaças à vida das populações, dos ecossistemas e do clima. Iniciativas como a plataforma de Gestão de Risco e Impactos de Queimadas e Incêndios Florestais (CEMADEN) podem contribuir para a redução destes riscos. Esta pesquisa teve como objetivo identificar o comportamento da ocorrência de focos de calor frente ao previsto no produto do CEMADEN. Identificou-se que a maior ocorrência de focos de calor, com 83,5%, está concentrada nos primeiros quatro trimestres de previsão (ASO, SON, OND e NDJ), e a diminuição nos níveis de alerta a partir da previsão de DJF. Esta análise contribui ainda para o entendimento da dinâmica do alerta e auxilia na priorização de municípios para ações de prevenção à queimadas.

1. Introdução

Os sistemas de alerta com antecedência figuram como um componente importante nas discussões para a redução do risco de desastres. Com a utilização desses sistemas é possível não apenas evitar a perda de vidas, como também o impacto econômico dos desastres [UNESCO, 2022]. Este entendimento junto a novas discussões e iniciativas de sistemas de alerta de desastres é observado no cenário mundial, em decorrência do Framework de Sendai para a Redução de Risco de Desastres, adotado em 2015 durante a Terceira Conferência Mundial das Nações Unidas.

As queimadas e os incêndios florestais estão inseridos nas categorias de ameaças que levam a impactos negativos na vida das populações, dos ecossistemas e do clima. Ameaças que na Amazônia apresentam-se acopladas ao desmatamento [ARAGÃO et

al., 2008]. Isso ocorre porque durante o processo de remoção da cobertura florestal, é feito uso do fogo para limpeza da biomassa, ao mesmo tempo que o desmatamento leva ao aumento da fragmentação florestal, que com o incremento da área de borda, há a intensificação dos seus efeitos, resultando na ocorrência de incêndios florestais [SILVA-JUNIOR *et al.*, 2018].

Neste contexto, o estado do Maranhão no Brasil chama atenção para a necessidade quanto à prevenção da ocorrência de novos incêndios, tendo em vista que é o segundo estado com maior nível de fragmentação florestal entre os estados da Amazônia brasileira [VEDOVATO *et al.*, 2016]. Assim, esta pesquisa teve como objetivo identificar o comportamento da ocorrência de focos de calor frente aos resultados do produto de previsão da probabilidade de fogo produzido pela Plataforma Gestão de Risco e Impactos de Queimadas e Incêndios Florestais - CEMADEN, de modo a validar os resultados do modelo de previsão com o comportamento observado para um município e enfatizar a importância de utilizar o sistema para evitar a ocorrência de grandes incêndios.

2. Material e Métodos

Mais de 115 mil km² da porção noroeste do estado do Maranhão é abrangida pelo bioma amazônico, distribuídos em 110 municípios que estão inseridos e/ou interseccionam o bioma (Figura 1). Estas áreas têm recebido um monitoramento para previsão da probabilidade de fogo desde agosto de 2021. A partir deste período, o Centro Nacional de Monitoramento e Alertas de Desastres Naturais (CEMADEN) iniciou sua disponibilização na plataforma de Gestão de Risco e Impactos de Queimadas e Incêndios Florestais (Figura 2).

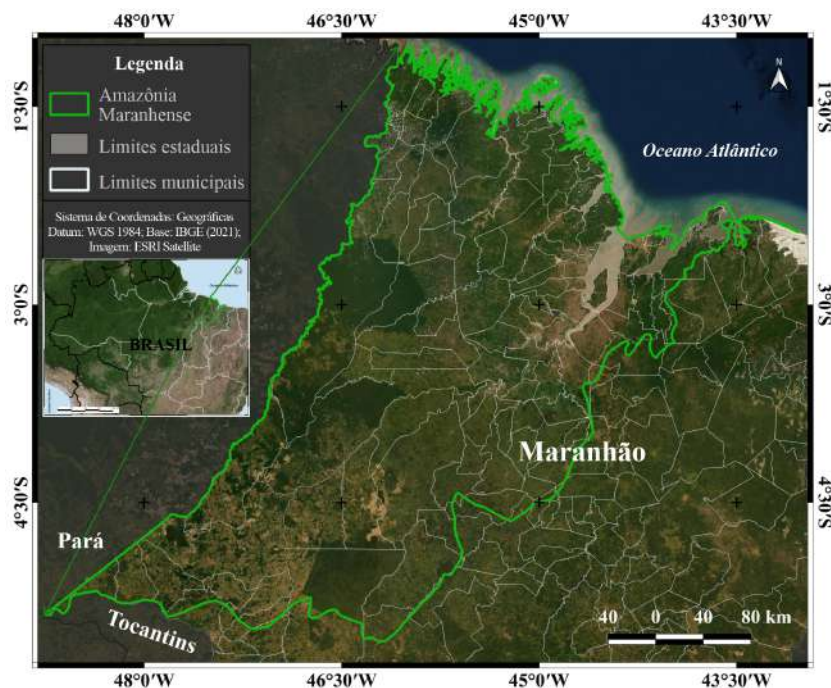


Figura 1. Mapa de localização da Amazônia Maranhense.

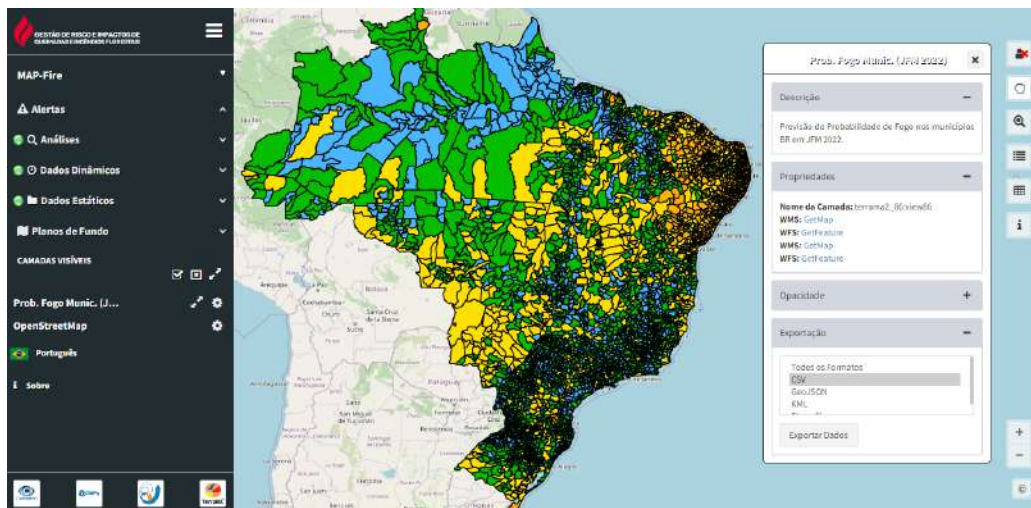


Figura 2. Plataforma de Gestão de Risco e Impactos de Queimadas e Incêndios Florestais [CEMADEN, 2022].

No total foram obtidos 11 períodos de previsão de probabilidade, contemplados pelo quantitativo de focos total para o trimestre, sendo obtido o comportamento para o intervalo temporal (Figura 3). Desta forma, foram validadas as ocorrências de focos de calor referente ao nível de alerta previsto para os municípios da Amazônia Maranhense.

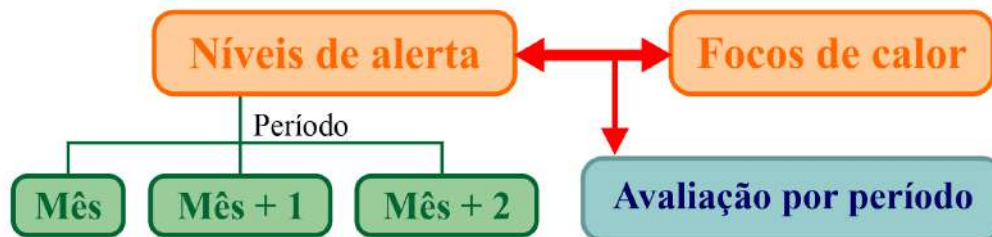


Figura 3. Avaliação da ocorrência de focos de calor x níveis de alerta.

3. Resultados

A maior ocorrência de focos de calor, com 83,5%, está concentrada nos primeiros quatro trimestres de previsão (ASO, SON, OND e NDJ), com picos de ocorrência em nível de Atenção e Alerta, de 19,5% em OND/2021 e 8% em SON/2021, respectivamente (Figura 4). Observa-se que a série apresenta uma diminuição na ocorrência de níveis de alerta mais altos, associados a maior ameaça ao fogo nos municípios, a partir da previsão de dezembro, DJF (**Figura 4**). Neste período é quando o estado de Observação passa a ser o maior registrado junto com a Baixa Probabilidade, com 40,7% e 23,9% do total de alertas emitidos. Nestes níveis a ocorrência total de focos de calor foi inferior a 18%, concentrados entre agosto/2021 e fevereiro/2022.

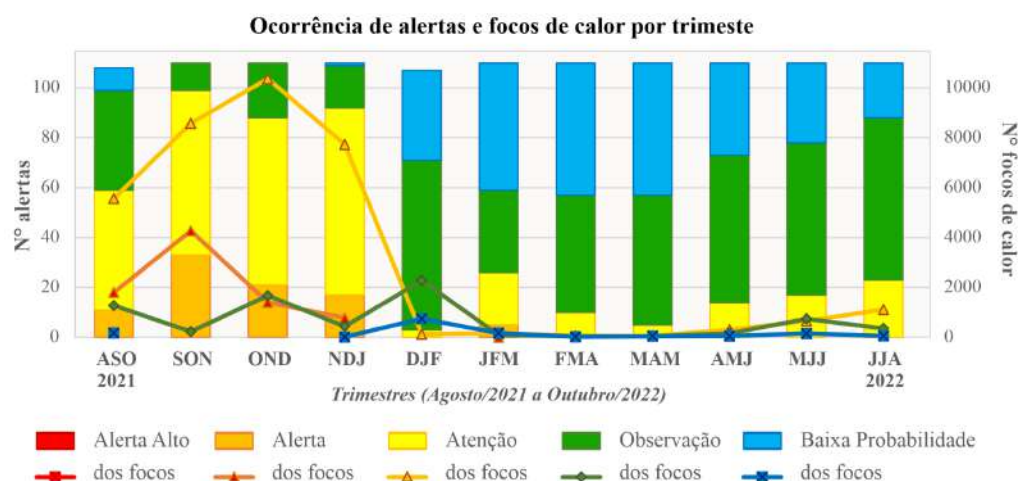


Figura 4. Ocorrência de alertas e focos de calor por trimestre.

4. Discussão

A dinâmica de ocorrência e dispersão das queimadas e incêndios florestais no Maranhão é modulada pelo comportamento dos padrões climáticos do estado (SILVA-JÚNIOR *et al.*, 2017). Estes eventos ocorrem principalmente durante o segundo semestre do ano (IMESC 2015), período em que inicia a estação de estiagem. Este comportamento é confirmado nos resultados encontrados nesta pesquisa, com o maior número de ocorrência de focos de calor registrados entre os meses de Agosto e Novembro, justamente para os primeiros trimestres de previsão da probabilidade de fogo em 2021 que registraram um alto número de focos no Maranhão.

Considerando este padrão já definido na literatura e observado neste trabalho, tanto na quantidade de ocorrências de focos de calor, quanto a maior quantidade de níveis de alertas mais altos no período de estiagem, é possível atestar o potencial das informações com antecedência. Logo, a utilização do produto pode confirmar sua função para o planejamento de atividades, ações de comunicação, priorização de áreas e mobilização de recursos e pessoal para a prevenção das queimadas e incêndios florestais. Uma vez que mesmo ao identificar um padrão de maior intensidade de níveis de alertas e confirmadas as ocorrências de focos, estas possuem uma distinção que permite um melhor gerenciamento por parte dos tomadores de decisão.

5. Conclusão

Neste trabalho foram avaliadas as ocorrências de focos de calor junto com os resultados do produto de previsão da probabilidade de fogo para os municípios da Amazônia Maranhense. Observou-se que nos primeiros quatro trimestre da série temporal (ASO, SON, OND, NDJ de 2021) a maioria dos municípios encontrava-se em situação de Atenção, chegando a quase 70% destes. Comportamento que influencia no elevado número de focos registrados na categoria.

A nível estadual, esta análise do primeiro ano de operação do produto de previsão contribui para o entendimento da sua dinâmica e a priorização de municípios

para ações de prevenção à queimadas e incêndios florestais no bioma no estado. Estas informações são fundamentais para o desenvolvimento de medidas para prevenir e mitigar o risco e os impactos do fogo enquanto ameaça. Ressalta-se, no entanto, a importância da inserção do contexto espacial nas análises futuras, para distinguir o comportamento da ocorrência de focos dentro das classes de alertas.

Referências

- Anderson, L. O.; Burton, C.; Reis, J. B. C.; Pessôa, A. C. M.; Bett, P.; Carvalho, N. S.; Silva-Junior, C. H. L.; Williams, K.; Selaya, G.; Armenteras, D. (2021) “An alert system for Seasonal Fire probability forecast for South American Protected Areas”, *Climate Resilience and Sustainability*.
- Aragão, L. E. O. C.; Malhi, Y.; Barbier, N.; Lima, A.; Shimabukuro, Y.; Anderson, L.; Saatchi, S. (2008) “Interactions between rainfall, deforestation and fires during recent years in the Brazilian Amazonia”, *Philosophical Transactions of the Royal Society B*, v. 363.
- CEMADEN - Centro Nacional de Monitoramento e Alertas de Desastres Naturais (2022) “Gestão de Risco e Impactos de Queimadas e Incêndios Florestais”, <http://terrama.cemaden.gov.br/griif/mapfire/monitor/>, setembro.
- IMESC - Instituto Maranhense de Estudos Socioeconômicos e Cartográficos (2015) “Análise da Incidência de Focos de Queimadas nas Terras Indígenas do Estado do Maranhão”, In:Relatório Técnico. São Luís. 25p.
- Silva-Junior, C.; Aragão, L.; Anderson, L.; Escada, I. (2017) “O papel da fragmentação e do efeito de borda na ocorrência e intensidade de incêndios florestais na Amazônia”. In: Simpósio Brasileiro de Sensoriamento, 18., Santos-SP, 2017. Anais... Santos-SP: INPE, p. 5952-5959
- Silva-Junior, C.; Aragão, L.; Fonseca, M.; Almeida, C.; Vedovato, L.; Anderson, L. (2018) “Deforestation-Induced Fragmentation Increases Forest Fire Occurrence in Central Brazilian Amazonia”. *Forests*.
- UNESCO - Organização das Nações Unidas para a Educação, a Ciência e a Cultura (2022) “Disaster Risk Reduction”, <https://en.unesco.org/disaster-risk-reduction/early-warning-systems>, setembro.
- Vedovato, L. B.; Fonseca, M. G.; Arai, E.; Anderson, L. O.; Aragão, L. E. O. C. (2016) “The Extent of 2014 Forest Fragmentation in the Brazilian Amazon”. *Reg. Environ. Change*, p. 2485–2490.

Agradecimentos

À FAPESP pelo Projeto MAP-Fire Maranhão (processo nº 2020/16457-3) e pela bolsa do subprojeto “Previsão da probabilidade de Incêndios florestais e balanço do Carbono para o Estado do Maranhão” (processo nº 2022/07714-8).

Uso de cubos de dados de observação da Terra no mapeamento de áreas queimadas

Mateus de Souza Macul¹, Wesley Campanharo¹, Fabiano Morelli², Gilberto Ribeiro de Queiroz¹, Karine Reis Ferreira¹

¹Divisão de Observação da Terra e Geoinformática (DIOTG)

²Divisão de Projeto Estratégico 4 - DIPE4

Instituto Nacional de Pesquisas Espaciais (INPE)

Caixa Postal: 515, Av. dos Astronautas, 1758 Jardim da Granja - SP - Brasil

{mateus.macul,wesley.campanharo,fabiano.morelli,gilberto.queiroz,
karine.ferreira}@inpe.br

Abstract. *Earth observation data cubes are multidimensional structures created from Analysis-Ready Data (ARD) of remote sensing images. This work presents the use of data cubes in the mapping of burnt area using the methodology of the product aq30m of the INPE's Queimadas Program. Tests were conducted in two different areas and at different scales in the Cerrado biome. The methodology was adapted to use the data cubes and the technologies developed by the Brazil Data Cube (BDC) project in R language. Subsequently, the results were compared with the original aq30m data and the advantages of using data cubes in burned area mapping were verified, enabling future improvements.*

Resumo. *Cubos de dados de observação da Terra são estruturas multidimensionais criadas a partir de dados prontos para análise (ARD - Analysis-Ready Data) de imagens de sensoriamento remoto. Este trabalho apresenta o uso de cubos de dados no mapeamento de área queimada utilizando a metodologia do produto aq30m do Programa Queimadas. Foram realizados testes em duas áreas distintas e em escalas diferentes no bioma Cerrado. A metodologia foi adaptada para utilizar os cubos de dados e as tecnologias desenvolvidas pelo projeto Brazil Data Cube (BDC) em linguagem R. Posteriormente, os resultados foram comparados com os dados originais aq30m e constatou-se as vantagens de utilizar cubos de dados no mapeamento de área queimada possibilitando futuros aprimoramentos.*

Introdução

Em 2008, a Organização das Nações Unidas para Alimentos e Agricultura (FAO) indicou área queimada como uma Variável Climática Essencial (do inglês, *Essential Climate Variable: ECV*) (SESSA; DOLMAN, 2008). Com isso, o mapeamento de área queimada, como uma das classes de mapeamentos de uso e cobertura do solo, é considerado como essencial para que países e agências internacionais monitorem, implementem e relatem questões relacionadas às mudanças climáticas. Dessa forma, é importante que tais observações atendam a requisitos mínimos para permitir entender as causas das mudanças climáticas, analisar os impactos potenciais, avaliar as opções de adaptação e possibilitar a caracterização de eventos extremos. Portanto, o painel do Sistema de Observação do Clima Global (do inglês, *Global Climate Observation System: GCOS*) estabeleceu que dados de áreas queimadas sejam de 30 metros de resolução espacial, com 24 horas de resolução temporal e com erros de comissão e omissão menores do que 15% (GCOS, 2016).

O Programa Queimadas do Instituto Nacional de Pesquisas Espaciais (INPE) realiza o monitoramento de focos de queimadas por meio de dados orbitais desde 1994. Mais recentemente, em 2015, o programa passou a realizar o mapeamento de área queimada com 30 m de resolução, resultando no produto denominado aq30m (INPE, 2019). Esse mapeamento utiliza como insumo imagens do programa Landsat da USGS (*United States Geological Survey*), o que exigiria baixar cada imagem individualmente diretamente do servidor da USGS. O armazenamento e pré-processamento das imagens tem demandado cada vez mais recursos computacionais.

O projeto *Brazil Data Cube* (BDC) do INPE gera e disponibiliza dados prontos para análise (ARD – *Analysis-Ready Data*) e cubos de dados multidimensionais de imagens de sensores a bordo dos satélites MODIS, Landsat 8, Sentinel-2 e CBERS 4. A partir desses cubos de observação da Terra é possível gerar mapas e padrões de mudança de uso e cobertura do solo, utilizando técnicas de aprendizado de máquina e análise de séries temporais de imagens (FERREIRA et al., 2020).

Portanto, este trabalho tem como intuito avaliar o processo de melhoria do mapeamento de área queimada utilizado pelo Programa Queimadas do INPE, para superar os desafios tecnológicos na preparação e armazenamento dos dados orbitais e atender os requisitos internacionais da geração de dados de monitoramento de queimadas. Para isso, este estudo propõe testar o método de geração do produto aq30m utilizando os cubos de dados de imagens Landsat gerados pelo projeto BDC.

Metodologia

O trabalho baseou-se na metodologia utilizada na produção do aq30m do Programa Queimadas, e a partir da sua análise implementar melhorias utilizando os dados e as ferramentas que o Projeto BDC propõe.

O método para produção do dado aq30m está descrito no trabalho realizado por MELCHIORI et al. (2014), BITTENCOURT et al. (2019), BITTENCOURT; MORELLI (2019) e BITTENCOURT et al. (2020). Em resumo, o processamento consiste em gerar polígonos de mudança de uso e cobertura que sejam possivelmente áreas queimadas a partir da diferença relativa de índices espectrais de imagens em tempos distintos (Figura 1).

A adaptação da metodologia do aq30m com os dados do BDC foi implementada em linguagem R (Figura 1), priorizando ferramentas e bibliotecas ligadas ao projeto e com capacidade de lidar com cubos de dados.

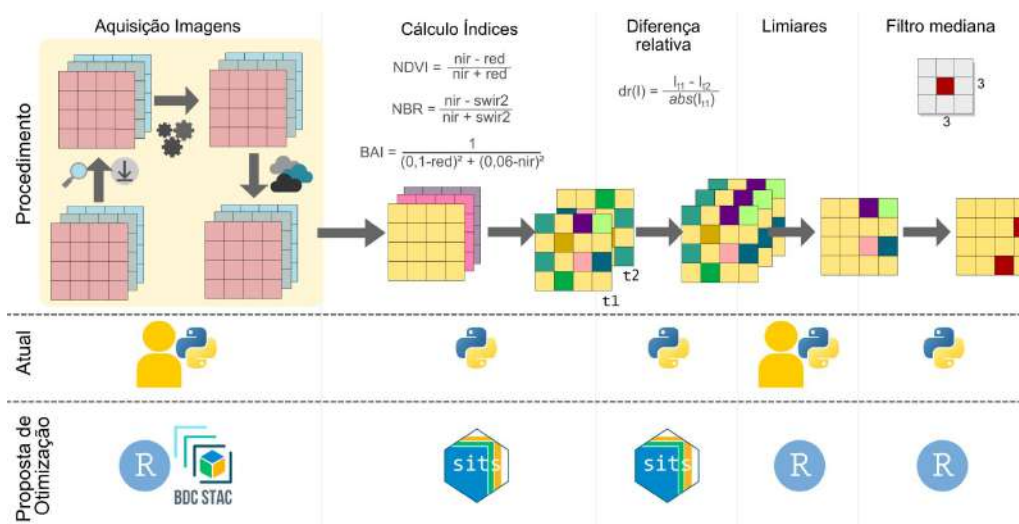


Figura 1 - Diagrama esquemático da metodologia.

Os dados utilizados são da coleção LC8_SR-1 do BDC que consiste em imagens corrigidas para reflectância de superfície oriundos dos sensores a bordo do satélite Landsat 8. A etapa seguinte é o cálculo dos índices espectrais. Na metodologia do aq30m, os índices utilizados são NDVI (*Normalized Difference Vegetation Index*), NBR (*Normalized Burn Ratio*) e BAI (*Burn Area Index*). Os índices são descritos por ROUSE, J. W. et al. (1974), KEY; BENSON (2006) e CHUVIECO; MARTÍN; PALACIOS (2002). Os índices foram calculados para dois instantes no tempo separadamente: t1 e t2.

A próxima etapa foi calcular as imagens de diferença relativa (*dr*), conforme mostrado na Figura 1. Após o cálculo da diferença relativa das imagens, aplicou-se um filtro mediana com grade 3 x 3 pixels. O intuito desta etapa é suavizar os valores discrepantes dos pixels. Por fim, para extrair os polígonos de mudança de cobertura do solo representando possivelmente a área queimada, o produto aq30m utiliza uma combinação lógica de fatiamento das imagens. Os limiares para o fatiamento são calculados para cada cena empiricamente por especialistas do Programa Queimadas. As imagens fatiadas foram combinadas com as máscaras de nuvem para o t1 e t2 e este resultado recebe uma suavização para manter somente as áreas com 6 pixels ou mais, conforme mostrado na Figura 2.

```
m.dnbr <- d_nbr >= 0.5
m.dndvi <- d_ndvi_rast >= 0.2
m.t2bai <- t2_bai >= 20
m.t1t2 <- (t1_nbr-t2_nbr) > 0.1
aq30m <- m.dnbr*m.dndvi*m.t2bai*m.t1t2*t1_fmask_rast*t2_fmask_rast
aq30m <- areaSieve(aq30m, thresh = (6*30*30), directions = 8)
```

Figura 2 - Código usado para fatiamento e álgebra dos rasters para geração das áreas queimadas

Área de Estudo

Para realizar os testes de aplicação das soluções desenvolvidas pelo BDC para mapeamento de área queimada foram identificadas 2 áreas de estudo no Bioma Cerrado (Figura 3) e estabeleceu-se o ano de 2020 como foco. Este bioma foi escolhido pois é a

mesma região de cobertura do dado aq30m do Programa Queimadas. As áreas de estudo foram escolhidas a partir de regiões com conhecimento prévio *in loco* e para realizar os testes em pequena e larga escala em relação a área. Portanto as áreas escolhidas foram a região do Assentamento Rural São Judas Tadeu, no município de Santa Rita do Tocantins, TO, na cena Landsat 223-068 (Figura 3a), e toda a região da cena Landsat 221-067 que corresponde a região do Parque Estadual do Jalapão (Figura 3b).

O Assentamento Rural São Judas Tadeu foi utilizado para realizar testes em pequena escala e é foco de estudo em tese de doutorado no programa de pós-graduação de Ciências do Sistema Terrestre do INPE.

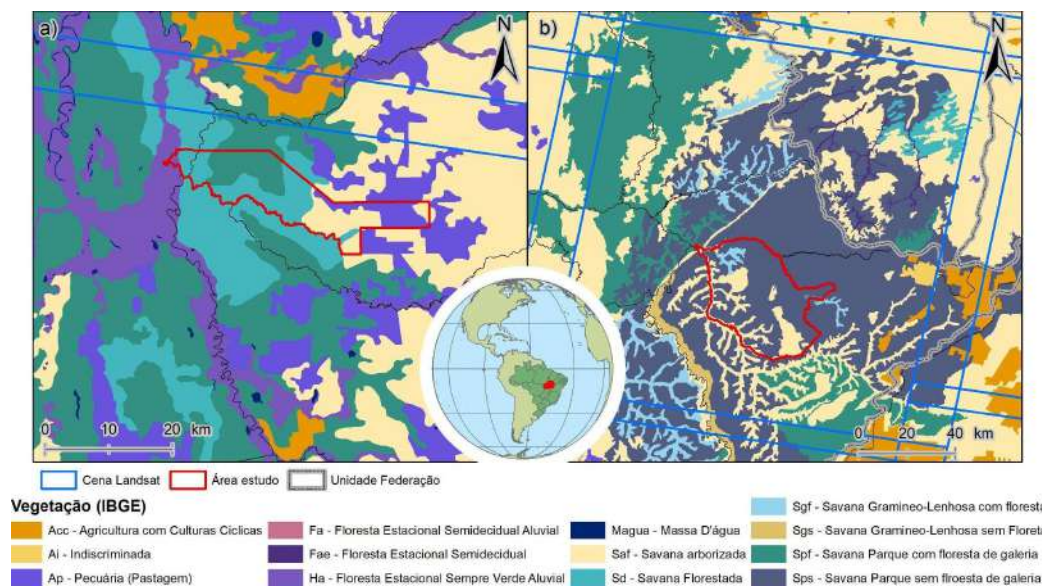


Figura 2 - Áreas de estudo a) região do Assentamento Rural São Judas Tadeu, cena Landsat 223-068; b) região do Parque Estadual do Jalapão, cena Landsat 221-067.

No ano de 2020, a cena Landsat 223-068 apresentou 17.646 focos ativos, 8.429 polígonos, totalizando 510.868 ha de área queimada. Esse dado foi obtido a partir de 6 imagens processadas no aq30m compreendendo o período de maio a setembro (INPE, 2019). É possível identificar também 5 fitofisionomias diferentes na região do assentamento rural. As imagens analisadas neste estudo foram capturadas em 31/08/2020 (t1) e 16/09/2020 (t2).

A região do Parque estadual do Jalapão é recorrentemente utilizada como área de testes no monitoramento pelo Programa Queimadas, incluindo diversos trabalhos de campo. Portanto, toda a área da Cena Landsat 221-067 foi designada para realizar os testes em maior escala.

Esta cena apresentou no ano de 2020 16.090 focos de fogo, 10.202 polígonos de aq30m, 665.195 ha de área queimada e foram processadas 10 imagens sem cobertura de nuvem no período de maio a novembro (INPE, 2019). Esta região também apresentou 6 fitofisionomias diferentes a partir de mapas do IBGE. As imagens foram capturadas em 17/08/2022 (t1) e 02/09/2020 (t2).

Resultados

O resultado da aplicação do BDC no método de geração de polígonos potencialmente de área queimada utilizado no produto aq30m pelo Programa Queimadas pode ser visualizado para as duas áreas de estudo na Figura 4. Nesta figura, também é possível comparar o resultado obtido com os dados originais produzidos pelo Programa Queimadas. Essa comparação permite avaliar o resultado deste trabalho, detalhada na Tabela 1.

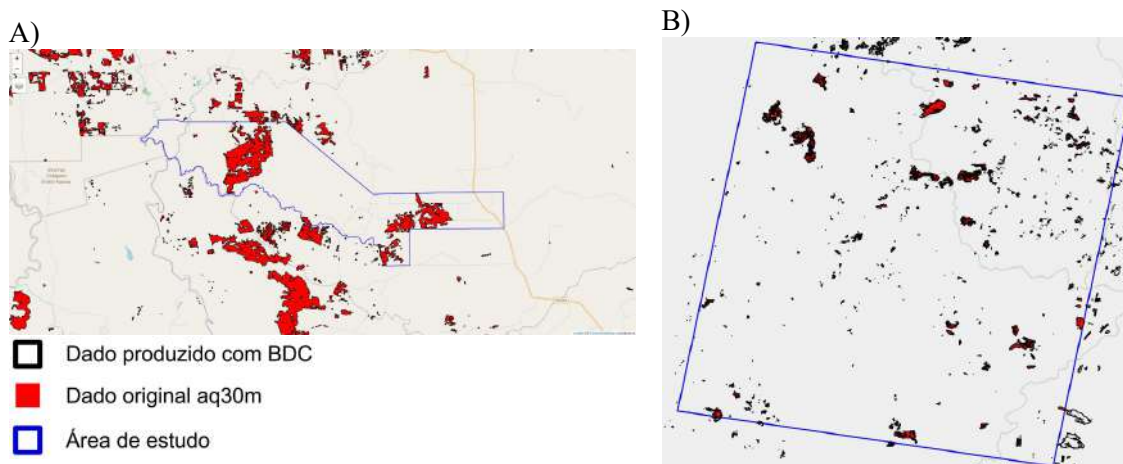


Figura 4 - Comparação dos polígonos gerados (preto) com os polígonos originalmente produzidos pelo Programa Queimadas (vermelho) para região do assentamento São Judas Tadeu (A) e para a cena Landsat 227-061 (B).

Tabela 1 - Comparação entre os resultados obtidos no mapeamento de área queimada por meio de cada metodologia empregada, considerando as duas áreas de estudo.

Metodologia	Assentamento São Judas Tadeu (A)		Cena Landsat 227-061 (B)	
	%	km ²	%	km ²
aq30m	0,72	1,00	0,68	3,54
Otimizada com BDC	1,68	2,32	0,74	3,86
Concordância	97,60	134,69	98,59	516,65
Total	100	138,00	100	524,05

Esse resultado mostrou que os polígonos produzidos a partir dos recursos desenvolvidos pelo BDC mantém em média uma concordância de 98,10%, uma discordância de aproximadamente 0.70% e um ganho de 1.21% comparado aos produzidos pelo aq30m do Programa Queimadas, considerando apenas o mapeamento de áreas queimadas.

Levando em consideração o ganho em tempo de processamento, os recursos que o BDC disponibiliza, otimiza principalmente a etapa de pré-processamento e armazenamento para preparação e disponibilização dos dados, uma vez que retira do especialista do Projeto Queimadas a necessidade de se realizar o download, a normalização das imagens, e a remoção de nuvens.

Outro mecanismo de otimização utilizado, foi o de processamento em paralelo atrelado ao processamento completo de um *tile*, oriundo dos recursos desenvolvidos do pacote SITS (*Satellite Image Time Series Analysis for Earth Observation Data Cube*) do BDC.

Este recurso foi utilizado em conjunto com o filtro de mediana, a etapa mais custosa computacionalmente dentre todo o processo. Com isso, o BDC se apresenta viável para integrar a produção do produto aq30m e possibilita expandir melhorias no mapeamento de área queimada nos dados oficiais.

Considerações Finais

Neste trabalho, foram realizadas importantes atividades de aplicação dos cubos de dados de observação da Terra e tecnologias desenvolvidas pelo projeto BDC para o aprimoramento no mapeamento de áreas queimadas. Os testes realizados em duas áreas diferentes evidenciaram a compatibilidade das tecnologias desenvolvidas pelo BDC para o mapeamento de mudança de uso e cobertura do solo usado na produção de dados de área queimada em conjunto com o Programa Queimadas do INPE. Portanto, futuros trabalhos de aplicação são promissores em testar novas melhorias na produção de mapas com o BDC, principalmente em parceria com outras iniciativas do instituto como o Programa Queimadas.

Por fim, este trabalho explorou e evidenciou aplicações somente com dados Landsat prontos para análise (ARD) gerados e disponibilizados pelo BDC. Entretanto, o BDC disponibiliza uma gama maior de dados e testes iniciais já começaram a ser realizados com outros dados, como o Sentinel. A tecnologia desenvolvida e mantida pelo BDC permite o acesso a grande quantidade de dados sem a necessidade de realizar downloads e pré-processamentos. Portanto, futuros trabalhos podem dar continuidade nesses estudos de aplicação.

References

- BITTENCOURT, O. O.; MORELLI, F. CLASSIFICAÇÃO DE ÁREAS QUEIMADAS POR MACHINE LEARNING. p. 1784–1787, 2019.
- BITTENCOURT, O. O.; MORELLI, F.; DOS SANTOS JÚNIOR, C. A.; SANTOS, R. Evaluating Classification Models in a Burned Areas' Detection Approach. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), v. 11622 LNCS, p. 577–591, 2019.
- BITTENCOURT, O. O.; MORELLI, F.; JÚNIOR, C. A. S.; SANTOS, R. An Approach to Classify Burned Areas Using Few Previously Validated Samples. p. 239–254, 2020.
- CHUVIECO, E.; MARTÍN, M. P.; PALACIOS, A. Assessment of different spectral indices in the red-near-infrared spectral domain for burned land discrimination. International Journal of Remote Sensing, v. 23, n. 23, p. 5103–5110, 2002.
- FERREIRA, K. R.; QUEIROZ, G. R.; VINHAS, L.; MARUJO, R. F. B.; SIMOES, R. E. O.; PICOLI, M. C. A.; CAMARA, G.; CARTAXO, R.; GOMES, V. C. F.; SANTOS, L. A.; SANCHEZ, A. H.; ARCANJO, J. S.; FRONZA, J. G.; NORONHA, C. A.; COSTA, R. W.; ZAGLIA, M. C.; ZIOTI, F.; KORTING, T. S.; SOARES, A. R.; CHAVES, M. E. D.; FONSECA, L. M. G. Earth observation data cubes for Brazil: Requirements, methodology and products. Remote Sensing, v. 12, n. 24, p. 1–19, 2020.
- GCOS. THE GLOBAL OBSERVING SYSTEM FOR CLIMATE: IMPLEMENTATION NEEDS. Guayaquil, Ecuador. 2016. Disponível em: <https://library.wmo.int/doc_num.php?explnum_id=3417>.
- INPE. Área Queimada — Programa Queimadas. 2019. Disponível em: <<https://queimadas.dgi.inpe.br/queimadas/portal/destaque/area-queimada>>. Acesso em: 26 abr. 2022
- KEY, C. H.; BENSON, N. C. Landscape Assessment: Ground measure of severity, the Composite Burn Index; and Remote sensing of severity, the Normalized Burn Ratio. 2006.
- MELCHIORI, A. E.; SETZER, A. W.; MORELLI, F.; LIBONATI, R.; CÂNDIDO, P. DE A.; JESÚS, S. C. DE. A Landsat-TM/OLI algorithm for burned areas in the Brazilian Cerrado: preliminary results. (D. X. Viegas, Ed.)In: Advances in Forest Fire research, VII International Conference on Forest Fire Research, 2014, Coimbra. Anais... Coimbra: Universidade de Coimbra, Portugal, 2014 Disponível em: <http://dx.doi.org/10.14195/978-989-26-0884-6_143>
- ROUSE, J. W., JR.; HAAS, R. H.; SCHELL, J. A.; DEERING, D. W. Monitoring vegetation systems in the Great Plains with ERTS. In: NASA. Goddard Space Flight Center 3d ERTS-1 Symp., Vol. 1, Sect. A, 1974, Anais... 1974
- SESSA, R.; DOLMAN, H. Terrestrial Essential Climate Variables for Climate Change Assessment, Mitigation and Adaptation. Roma. 2008. Disponível em: <<http://scholar.google.com/scholar?hl=en&btnG=Search&q=intitle:Terrestrial+Essential+Climate+Variables+for+Climate+Change+Assessment,+Mitigation+and+Adaptation#6>>.

Mapeamento da susceptibilidade á incêndios através de Sensoriamento Remoto no município de Cáceres

Felipe N. Souza¹, Rogério G. Negri¹, Vinicius L. S. Gino¹, José V. O. Simões¹

¹Instituto de Ciência e Tecnologia – ICT
Universidade Estadual Paulista “Júlio de Mesquita Filho” – UNESP
12247-004 – São José dos Campos – SP, Brasil

{fn.souza, rogerio.negri, vinicius.gino, jose.orlandi}@unesp.br

Abstract. *Recently, fire events recorded in Brazil have motivated the application of Remote Sensing and Machine Learning techniques in order to enable the monitoring and prevention of the impacts associated with these events. In this way, the present project aims to map the susceptibility to fires in forest regions through the application of Logistic Regression, built on the basis of the extraction of statistical descriptors under spectral indices.*

Resumo. *Recentemente os eventos de queimada registrados no Brasil tem motivado a aplicação de técnicas de Sensoriamento Remoto e Aprendizado de Máquina a fim de possibilitar o monitoramento e a prevenção dos impactos associados à estes eventos. Tendo isso em vista, o presente projeto tem como objetivo mapear a susceptibilidade à incêndios em regiões de floresta através da aplicação de Regressão Logística, construídos com base na extração de descritores estatísticos sob índices espectrais.*

1. Introdução

Nos últimos 20 anos o número de focos de incêndio registrados no Brasil aumentou significativamente, em decorrência da grilagem de terras e do uso do fogo na agricultura e pecuária, provocando uma série de impactos ambientais observados ao decorrer do tempo nas regiões afetadas por estes eventos [de Oliveira-Júnior et al. 2020]. Além dos impactos causados à biodiversidade local provocados por esse tipo de evento, a emissão de gases do efeito estufa e de material particulado em decorrência das queimadas, a intensificação de doenças respiratórias causadas por esse tipo de emissão pode afetar de forma direta comunidades que vivem ao redor desses eventos [de Oliveira-Júnior et al. 2020, Caúla et al. 2015].

Neste contexto, o monitoramento e a análise de eventos de queimada tornam-se imprescindíveis no entendimento das dinâmicas de alteração do uso do solo, além das condições climáticas e ecológicas. O Sensoriamento Remoto, com foco na problemática exposta anteriormente, pode representar uma ferramenta de extrema importância no monitoramento de queimadas, possibilitando que a análise proposta seja desenvolvida de forma multitemporal e de forma confiável [Granemann and Carneiro 2009]. No âmbito do desenvolvimento de ferramentas de monitoramento, as técnicas de classificação e regressão de dados, além da extração de índices espectrais através de ferramentas de sensoriamento remoto, quando aliadas, viabilizam a extração de conhecimento da superfície terrestre de forma automatizada.

Diante desta considerações, o presente trabalho tem como objetivo mapear a susceptibilidade à incêndios em parte do município de Cáceres, localizado no estado do Mato Grosso, utilizando modelos de regressão logística construídos a partir de registros históricos de queimada.

2. Área de estudo e dados disponíveis

A área de estudo utilizada (Figura 1) compreende o município de Cáceres, estado do Mato-Grosso, que abrange aproximadamente 24,6 mil quilômetros quadrados de extensão. Este município compõe a região Centro-Sul Mato-grossense e é dividido em três biomas predominantes: Amazônia, Cerrado e o Pantanal. Dentre os três biomas citados anteriormente, o Pantanal é o bioma mais expressivo no município de Cáceres, ocupando cerca de 51% de sua área total [IBGE 2020].

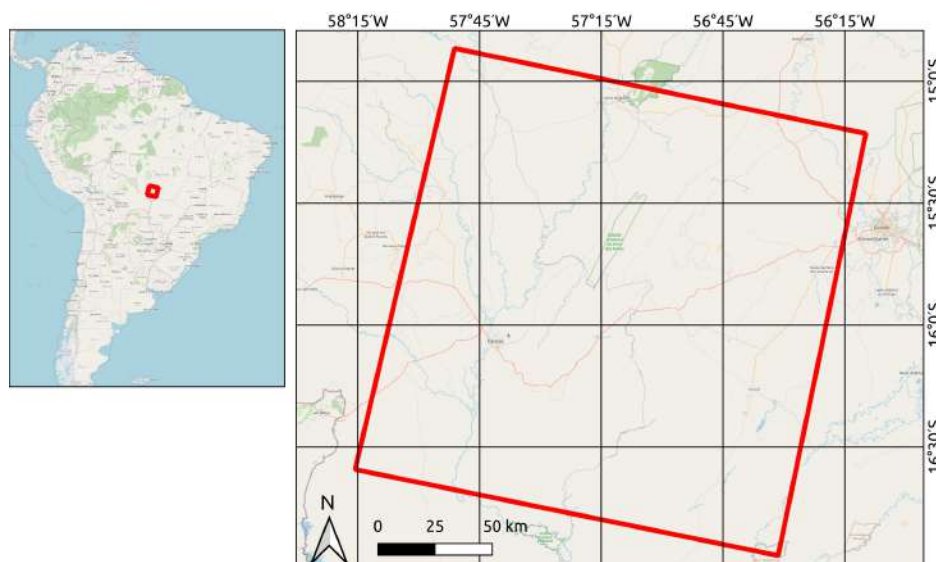


Figura 1. Área de estudo – região do município de Cáceres

Como base de dados, foram utilizadas séries de imagens tomadas pelo sensor *Operational Land Imager* (OLI), acoplado ao satélite Landsat-8. Índices espectrais foram computados a partir das imagens que compõem a série empregada. Os índices considerados foram: Índice de Vegetação por Diferença Normalizada (NDVI) [Rouse et al. 1974]; Índice de Água por Diferença Normalizada (NDWI) [Gao 1996]; e Índice de Queimada por Razão Normalizada (NBR) [Roy et al. 2006]. As imagens foram obtidas por meio da Interface de Programação de Aplicação do Google Earth Engine (API-GEE) [Gorelick 2013] via Python 3.8. A manipulação e construção dos conjuntos de dados utilizados na pesquisa foram realizadas com apoio das bibliotecas *Numpy* e *Pandas*. A aplicação de modelos de regressão foi realizada através das bibliotecas *Scikit-Learn* e *SciPy*.

3. Mapeamento da susceptibilidade à incêndios

A modelagem da susceptibilidade à incêndios proposta parte inicialmente, através do produto *MODIS Burned Area* [Giglio 2015], da identificação das regiões afetadas, ou não,

por incêndios dentro do período estipulado e na área de interesse apresentada na Seção 2. Posteriormente, com base na identificação dos períodos em que houveram queimadas na área de interesse, obtém-se uma coleção de imagens sobre a mesma localização e com cobertura de nuvens inferior a 5% dentro de um período de Q dias anteriores ao evento de queimada registrado pelo *MODIS Burned Area*. No presente estudo, o período considerado foi de 30 dias anteriores à cada um dos eventos de queimada registrados pelo *MODIS Burned Area* entre os anos de 2013 e 2021. Em seguida, foram extraídos os índices NDVI, NDWI e NBR a fim de caracterizar o uso e ocupação do solo na área de estudo e observar o seu comportamento em relação aos eventos de fogo registrados. O mesmo processo é repetido em locais não afetados por fogo, proporcionando assim um conjunto de exemplos complementar à ocorrência de queimada.

Em posse dos índices espectrais e tendo identificado regiões afetadas, ou não, por fogo, foi construído um conjunto de dados com descritores estatísticos (e.g., média e mediana), os quais são utilizados como meio de expressar as dinâmicas locais do solo. Por fim, o conjunto de dados obtido foi utilizado no treinamento de um modelo de regressão logística cuja aplicação sobre imagens em instantes recentes produz uma representação da susceptibilidade à futuros incêndios. A análise e avaliação dos resultados será realizada mediante a construção de histogramas e a aplicação do teste de Mann-Whitney [McKnight and Najab 2010].

4. Resultados

A Figura 2 exibe o mapeamento obtido segundo o procedimento discutido na Seção 3. Tal resultado é expresso em valores no conjunto $[0, 1]$, portanto, foram definidas, conforme apresentado na Tabela 1, diferentes classes de suscetibilidade à incêndios.

Tabela 1. Classes de suscetibilidade.

Suscetibilidade à incêndios	Classes de suscetibilidade
$[0, 0.2)$	Muito baixo
$[0.2, 0.4)$	Baixo
$[0.4, 0.6)$	Média
$[0.6, 0.8)$	Alto
$[0.8, 1]$	Muito alto

A fim de verificar a eficácia do procedimento proposto, foram construídos dois histogramas em relação à frequência com que diferentes probabilidades de incêndio foram identificadas pelo modelo. É observado na Figura 3 que, nas regiões afetadas por fogo, o modelo proposto apontou maior suscetibilidade à incêndios em contraponto às regiões não afetadas por fogo.

Observa-se através do comportamento registrado nos histogramas que, nas regiões que foram afetadas por fogo, o modelo proposto atribuiu uma suscetibilidade à incêndio maior que nas regiões que não foram afetadas por fogo. A aplicação do teste de Mann-Whitney nos dois conjuntos de dados obtidos comprovou a hipótese levantada, pois retornou $p - valor < 10^{-6}$, indicando que as duas amostras analisadas diferem estatisticamente, evidenciando a diferença entre o resultado retornado pelo modelo em dois grupos distintos.

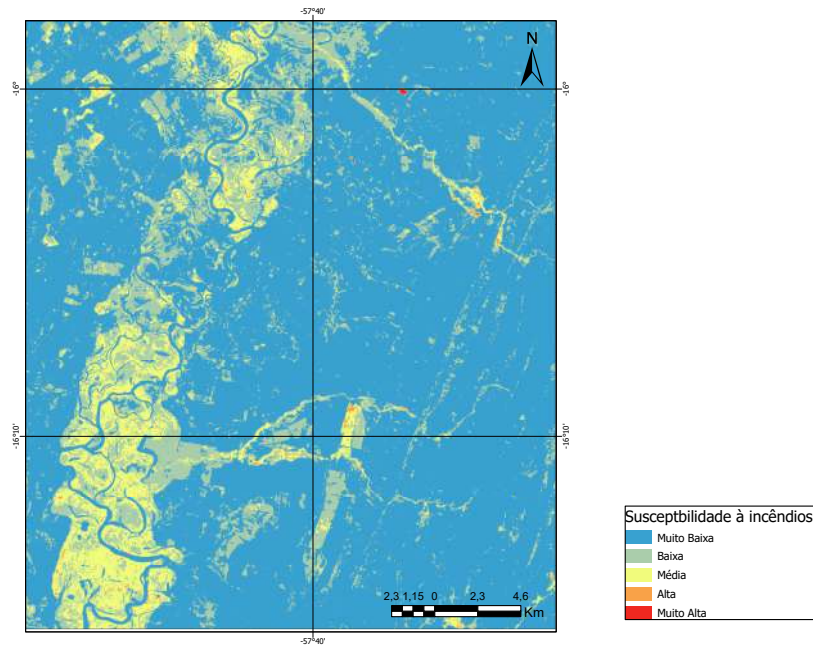


Figura 2. Mapeamento da Susceptibilidade à Incêndios na área de estudo

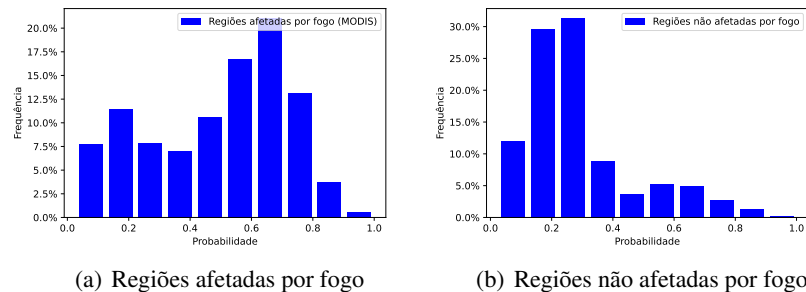


Figura 3. Histogramas da susceptibilidade à incêndios em regiões afetadas e não afetadas por fogo

5. Conclusão e perspectivas futuras

Os resultados obtidos com a aplicação do procedimento proposto mostrou-se consistentes quando analisados segundo o teste de Mann-Whitney. Segundo uma análise qualitativa, conforme definido na Figura 2, é possível identificar que as regiões referentes às amostras de fogo apresentaram alta probabilidade/susceptibilidade à incêndios, corroborando assim com a validação do modelo aplicado, porém, observa-se no mapeamento obtido que alguns pixels isolados apresentaram susceptibilidade à incêndio muito elevada, interferindo no resultado final do mapeamento e prejudicando parcialmente a análise visual e estatística proposta.

Apesar do resultado obtido ter sido satisfatório, espera-se aplicar e avaliar os resultados segundo diferentes modelos de regressão, permitindo avaliar assim o modelo mais adequado ao sistema proposto. Espera-se também integrar o resultado obtido ao algoritmo

Iterated Conditional Modes (ICM) [Weinman 2008], que por sua vez proporcionará modelos de regressão mais aderentes em termos espaciais, e logo, com maior precisão em relação à definição da susceptibilidade à incêndios na área de estudo.

6. Agradecimentos

Os autores agradecem à FAPESP (Procs. 2020/14664-1 e 2021/01305-6) pelo apoio financeiro.

Referências

- Caúla, R. H., Oliveira-Júnior, J. F., Lyra, G. B., Delgado, R. C., and Heilbron Filho, P. F. L. (2015). Overview of fire foci causes and locations in brazil based on meteorological satellite data from 1998 to 2011. *Environmental Earth Sciences*, 74(2):1497–1508.
- de Oliveira-Júnior, J. F., Teodoro, P. E., da Silva Junior, C. A., Baio, F. H. R., Gava, R., Capristo-Silva, G. F., de Gois, G., Correia Filho, W. L. F., Lima, M., de Barros Santiago, D., Freitas, W. K., dos Santos, P. J., and da Silva Costa, M. (2020). Fire foci related to rainfall and biomes of the state of mato grosso do sul, brazil. *Agricultural and Forest Meteorology*, 282-283:107861.
- Gao, B.-C. (1996). Ndwí? a normalized difference water index for remote sensing of vegetation liquid water from space. *Remote sensing of environment*, 58(3):257–266.
- Giglio, L., J. C.-B. L. R. D. (2015). Mcd64a1 modis/terra+aqua burned area monthly 13 global 500m sin grid v006. *NASA EOSDIS Land Processes DAAC*.
- Gorelick, N. (2013). Google earth engine. In *EGU general assembly conference abstracts*, volume 15, page 11997. American Geophysical Union Vienna, Austria.
- Granemann, D. C. and Carneiro, G. L. (2009). Monitoramento de focos de incêndio e áreas queimadas com a utilização de imagens de sensoriamento remoto. *Revista de Engenharia e Tecnologia*, 1(1):Páginas–55.
- IBGE (2020). Instituto brasileiro de geografia e estatística.
- McKnight, P. E. and Najab, J. (2010). Mann-whitney u test. *The Corsini encyclopedia of psychology*, pages 1–1.
- Rouse, J., Haas, R. H., Schell, J. A., Deering, D. W., et al. (1974). Monitoring vegetation systems in the great plains with erts. *NASA special publication*, 351(1974):309.
- Roy, D., Boschetti, L., and Trigg, S. (2006). Remote sensing of fire severity: assessing the performance of the normalized burn ratio. *IEEE Geoscience and Remote Sensing Letters*, 3(1):112–116.
- Weinman, J. (2008). A brief introduction to icm.

Ionospheric Scintillation Awareness for GPS and Other Constellations by Probability Maps

André Ricardo Fazanaro Martinon, Stephan Stephany, Eurico Rodrigues de Paula

Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758, CEP 12227-010, São José dos Campos SP, Brazil
{andre.martinon, stephan.stephany, eurico.paula}@inpe.br

Abstract. *Ionospheric scintillation may disturb radio-frequency signals and other electromagnetic waves. In the case of Global Navigation Satellite Systems, strong scintillation can cause loss-of-lock between satellite and ground receiver, and thus positioning errors. Remote sensing satellites can also be affected, generating distorted images. Scintillation maps covering the globe are available, but refer to the vertical projection of the scintillation, while satellite signals are emitted or received at an angle in the majority of the cases. Therefore, this work proposes the use of probability maps, generated from standard scintillation maps, in order to assess the risk of undergoing scintillation according to the satellite orbit.*

1. Introduction

This work proposes the innovative scintillation probability map for evaluating if the satellite-ground station links of a Global Navigation Satellite System (GNSS) like GPS can be affected by strong ionospheric scintillation, causing loss-of-lock or positioning error. Scintillation maps covering the globe are normally available in near real time, yielding the vertical projection of the amplitude scintillation index S4 at every grid point with a given time resolution. These maps are intended to be a representation of scintillation in the ionosphere. However, considering the scintillation map for a particular time interval, a GNSS receiver at a given latitude and longitude will not experience scintillation effects corresponding to its location, since the azimuth and elevation angles of each satellite-receiver link must be taken into account. Therefore, such scintillation probability maps are proposed, which provide for every grid point a probability of a receiver undergoing scintillation in function of the average positions of the set of GNSS satellites within reach in the considered time interval. Satellite positions can be promptly estimated from IGS (International GNSS Service) “Precise Satellites Orbits” data, describing the satellite orbits. Consequently, neighboring scintillation occurrences are taken into account in the probability map, providing to GNSS users a better insight about the possibility of being affected by strong scintillation near or at their position.

These newly proposed probability maps are made from different scintillation maps covering most of South America, which are more precise than standard maps due to a better interpolation algorithm and pre-processing schemes [Martinon et al., 2022], but the latter could also be employed. Typically, the S4 index is estimated from radio signals acquired at 50 Hz during a 1-minute period, yielding a S4 value given by the standard deviation of the signal intensity normalized by its mean [Van Dierendonck et al., 1993]. These values are calculated for each GNSS satellite-station link using the GNSS constellations currently available (GPS and others) providing compatibility even between stations that use different receivers [de Paula et al., 2021]. The present work employs only the GPS constellation to produce the scintillation and then the probability maps. It is expected that the proposed ionospheric scintillation probability maps will provide warnings for users of GNSS-based applications, including aerial navigation, ADS-B (Automatic Dependent Surveillance–Broadcast), takeoff and landing

procedures, precision agriculture, and positioning of offshore oil prospecting platforms. In addition, these maps can be assimilated by ionospheric models and/or used in scintillation research, besides being included as part of the monitoring services provided by the Brazilian Airspace Control System (SISCEAB).

The probability maps presented here are specific for GNSS applications, but the same approach can be easily adapted for remote sensing satellites, by taking into account the scan angle to nadir for each swath of the considered onboard sensor. The same scintillation map may then be employed to generate customized probability maps, according to the remote sensing satellite orbit and scan angle. Besides radio-frequency, some other electromagnetic waves may be affected by scintillation, as in the case of SAR (Synthetic Aperture Radar). On the other hand, distortions in SAR images caused by scintillation may allow to identify ionospheric bubbles [Sato et al., 2021], large structures of depleted plasma that contain high density gradients associated with the occurrence of scintillation.

2. Ionospheric Scintillation Maps Generation

The ionospheric scintillation maps used in this work were generated using S4 index data of GPS receivers covering most of South America, and provided by the GNSS stations of two networks, LISN (Low-Latitude Ionospheric Sensor Network) [Valladares and Chau 2012] and INCT GNSS-NavAer [Vani et al., 2017]. The data collected by each receiver of the monitoring station includes the GPS satellite positions (azimuth and elevation) and the corresponding ionospheric scintillation indexes S4.

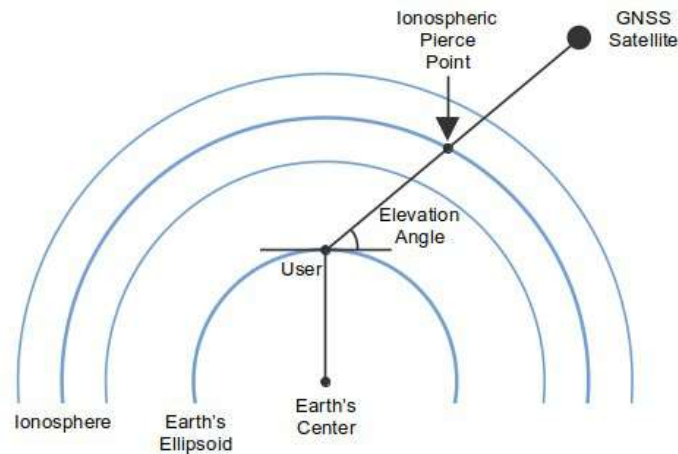


Figure 1. Geometry of a satellite-ground station line-of-sight depicting the Ionospheric Pierce Point

The making of a scintillation map covering a time interval (here, 15-minute) requires some assumptions/steps: (i) the ionosphere is modeled as a thin spherical shell over the Earth at the mean altitude of 350 km, which corresponds to the maximum observed TEC (Total Electron Content); (ii) for each satellite-receiver link, the position of a Ionospheric Pierce Point (IPP, shown in Figure 1) can be obtained as the intersection of the line-of-sight receiver-satellite with the ionosphere, using the receiver coordinates and azimuth and elevation angles of the satellite [Prol et al., 2017]; (iii) each line-of-sight satellite-receiver is a slant path, i.e. the GNSS signal traverses a longer path inside ionospheric irregularities than for a vertical path, depending on the elevation angle; a way to normalize S4 values of observations with different elevation angles is to project each value to the vertical direction, assuming a line-of-sight of 90° at the IPP [Spogli et al., 2009 and 2013]; (iv) considering a $1.0^\circ \times 1.0^\circ$ resolution grid, the IPPs are grouped into the corresponding cells, and the S4 value for each cell is given

by the corresponding S4 values of the IPPs reduced by the maximum value, mean value or mean value of the 3rd quartile; (v) the interpolation of the cell S4 values yields the scintillation map, covering in this work, longitudes in the range -78.0° to -30.0° , and latitudes -39.0° to 9.0° , both with a 0.25° resolution. Differently from standard methods like Inverse Distance Weighting [Rezende et al., 2007], interpolation is performed using Gaussian Process Regression with a specific kernel, and adopting the Haversine distance as the distance metric. Figure 2 shows the resulting 15-minute scintillation map for October 10, 2014 at 00:31 UTC time, using S4 values projected to the vertical, and reducing the cell grouped IPPs by the mean of the 3rd quartile of their S4 values.

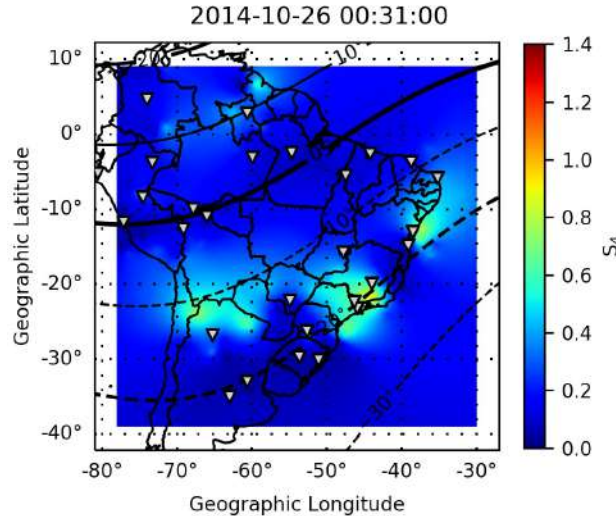


Figure 2. Example of a scintillation map, where color scale refers to S4 values. Curved lines depict magnetic latitudes, and triangles show GPS stations.

3. Proposed Methodology for Scintillation Probability Maps

This work proposes for the first time the making of scintillation probability maps, which are derived from the scintillation maps for the corresponding area and time interval. Such a map provides for a receiver location the probability of undergoing scintillation, specifically strong scintillation (S4 above 0.70), though it can be extended for any other scintillation level, like null [0.00-0.15], weak (0.15-0.30], or moderate (0.30-0.70]. A GPS user may consult scintillation probability maps, which can be generated in real time, in order to mitigate consequences of a possible loss of lock during position assessment or navigation.

This probability is calculated for every grid point of the map retrieving the S4 values of nearby points in the scintillation map, defined using the positions of the set of in-range GPS satellites with elevation angle higher than 30° , and thus take into account the influence of the elevation and azimuthal angles of the line-of-sight satellite-station for a hypothetical station at the grid point. Each GPS satellite position can be calculated by the available IGS satellite orbit data. The hypothetical GPS receiver is assumed as locking up to all satellites in its vicinity, thus defining a set of surrounding IPPs, each one with a S4 value obtained from the corresponding scintillation map interpolating the S4 values of the four closest neighboring grid points. The probability at the grid point is then simply given by the ratio of IPPs with strong scintillation (or any other level) and the total number of IPPs. The latter number varies according to the grid point latitude, decreasing as the latitude increases. Figure 3 shows the resulting 15-minute strong scintillation probability map for October 10, 2014 at 00:31 UTC time derived from the corresponding scintillation map shown in Figure 2.

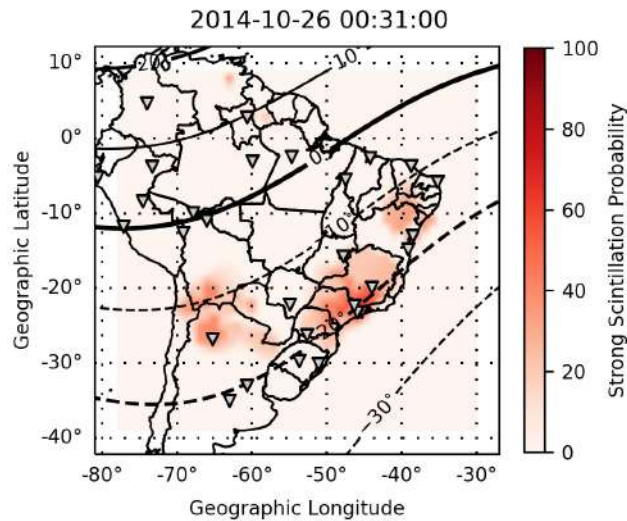


Figure 3. Corresponding strong scintillation probability map, where color scale refers to strong scintillation probability. Curved lines depict magnetic latitudes, and triangles show GPS stations

These scintillation probability maps are intended to evaluate for a user with a GPS receiver on ground or in a vehicle (car, plane, boat, etc.) the risk of suffering a loss-of-lock that may preclude the receiver from acquiring an updated or correct geographic position.

4. Results and Discussion

The proposed probability map approach was evaluated for a set of 40 (15-minute) intervals from November, 2013 to November, 2014, which were selected as presenting strong scintillation over the Brazilian territory. Previously, the corresponding 40 scintillation maps were made using an also newly proposed methodology for pre-processing and interpolation, employed here. Six pre-processing options were tested, combining operators to reduce S4 values for grouped IPPs (using mean, max, or 3rd quartile), and the use of slant or vertically-projected S4 values. The scintillation probability maps were then derived for the strong scintillation level and, for each GPS station, the specific probability was obtained by interpolating the closest four grid points of the probability map. Several error metrics were calculated using the real set of S4 values from all the IPPs of all GPS stations. Considering the four scintillation classes described in the beginning of the former section, a simple count of the number of IPPs of each scintillation class for each station divided by the total of IPPs yields 4 real station-linked ratios for the complete data set, but only the strong level one matters here, being named true strong-class ratio for each station, and composing a set of true strong-class ratios.

In order to evaluate the use of standard scintillation maps, another set of strong-class ratios was calculated for all stations, for the same real IPP positions for each of the 40 intervals, but using the S4 values interpolated directly from the scintillation map, instead of the set of real values. This approximate set of strong-class ratios is named baseline set, likely rendering the lower possible error compared to the true set, also for each pre-processing option employed to produce the scintillation maps. Please note that the scintillation map will provide a GNSS station with a S4 value for its position that is certainly far less accurate than the corresponding one in the baseline set.

The proposed probability map for the considered time interval allows to interpolate the four nearest neighboring grid points to obtain a probability at a given

station for any scintillation class. Considering the strong scintillation class, such probability is equivalent to the true and baseline sets. Finally, the probability and the baseline sets are compared to the corresponding true set for each station. Seven error metrics were employed in the comparison: overall Average (AVG), overall Mean Absolute Error (MAE), overall Root Mean Square Error (RMSE), 40-Mean of Minimum Errors (MIN), 40-Mean of Maximum Errors (MAX), overall Standard Deviation of Absolute Error (STD) and 40-Mean of Pearson's Correlation Coefficient (CORR).

Table 1 shows the error results for the baseline set and for the corresponding values of the proposed probability maps, taking as reference the true set. The best pre-processing options for the baseline set were obtained using the S4 vertical projection and reduction by the mean of the 3rd quartile, while the worst results arose for both slant and vertical S4 values with reduction by the plain mean. Considering the proposed probability maps, the best option was also the S4 vertical projection and reduction by the mean of the 3rd quartile. As expected, errors for the baseline set were lower than for the proposed probability maps, but the latter errors were quite acceptable for a future operational use.

Table 1. Error evaluation for the baseline set of strong-class ratios and the proposed probability maps

ERRORS FOR THE BASELINE SET							
AVG	MAE	RMSE	MIN	MAX	STD	CORR	OPTIONS
5.79	6.07	13.11	-2.39	37.62	11.62	0.87	slant_max
-2.03	2.35	5.27	-15.08	3.03	4.71	0.84	slant_mean
3.08	3.75	8.49	-4.58	25.02	7.62	0.92	slant_quantile
5.42	5.75	12.5	-2.56	35.92	11.1	0.87	vertical_max
-2.14	2.41	5.43	-15.53	2.36	4.87	0.83	vertical_mean
2.8	3.51	7.96	-4.83	23.62	7.15	0.91	vertical_quantile
ERRORS FOR THE PROBABILITY MAP							
AVG	MAE	RMSE	MIN	MAX	STD	CORR	OPTIONS
12.2	12.25	21.64	-0.67	55.61	17.83	0.78	slant_max
2.93	3.97	8.39	-5.74	24.32	7.39	0.84	slant_mean
9.41	9.5	17.45	-0.98	46.88	14.64	0.81	slant_quantile
5.27	5.77	12.13	-3.87	32.51	10.67	0.84	vertical_max
-2.44	3.05	6.52	-17.43	4.79	5.76	0.74	vertical_mean
2.72	3.72	8.21	-6.09	22.46	7.31	0.87	vertical_quantile

The innovative scintillation probability maps were evaluated using an extensive amount of real S4 index data obtaining low errors according to some metrics, showing the potential of the approach as a real warning tool for strong scintillation in GNSS applications, and also remote sensing and telecommunication satellites in general that employ electromagnetic bands that can be affected by scintillation.

Further works devise the making of scintillation and scintillation probability maps employing multi-constellation GNSS data, instead of only GPS data. Additionally, a specific pre-processing for S4 index missing data due to station technical issues or

loss-of-lock is also intended. Preliminary evaluations already show that real time data acquisition and processing for scintillation and scintillation probability maps making for operational purposes is feasible using standard computational and network resources.

References

- de Paula, E. R., Martinon, A. R. F., Moraes, A. O., Carrano, C., Neto, A. C., Doherty, P., et al. (2021). Performance of 6 different global navigation satellite system receivers at low latitude under moderate and strong scintillation. *Earth and Space Science*, 8, e2020EA001314. <https://doi.org/10.1029/2020EA001314>
- Martinon, A. R. F., Stephany, S. and de Paula, E. R. (2022). Challenges in Real-Time Generation of Scintillation Index Maps. Submitted to the XII Colóquio Brasileiro de Ciências Geodésicas/V Simpósio Brasileiro de Geomática, November 2022, Curitiba/PR - Brazil.
- Prol, F. S., Camargo, P. O. and Muella, M. T. A. H. (2017). Comparative Study of Methods for Calculating Ionospheric Points and Describing the GNSS Signal Path. *Boletim de Ciências Geodésicas*. 2017, v. 23, n. 4, pp. 669-683, ISSN 1982-2170. <https://doi.org/10.1590/S1982-21702017000400044>.
- Rezende, L. F. C., de Paula, E. R., Kantor, I. J. and Kintner, P. M. (2007). Mapping and survey of plasma bubbles over Brazilian territory. *The Journal of Navigation*, 60(1), 69– 81.
- Sato, H., Kim, J. S., Otsuka, Y., Wrasse, C. M., Rodrigues de Paula, E., and Rodrigues de Souza, J. (2021). L-band Synthetic Aperture Radar observation of ionospheric density irregularities at equatorial plasma depletion region. *Geophysical Research Letters*, 48, e2021GL093541. <https://doi.org/10.1029/2021GL093541>
- Spogli, L., Alfonsi, L., de Franceschi, G., Romano, V., Aquino, M. H. O. and Dodson, A. (2009). Climatology of GPS ionospheric scintillations over high and mid-latitude European regions. *Ann. Geophys* 27, 3429–3437.
- Spogli, L., Alfonsi, L., Romano, V., de Franceschi, G., Monico, G., Shimabukuro, M., Bougar, B. and Aquino M. (2013). Assessing the GNSS scintillation climate over Brazil under increasing solar activity. *Journal of Atmospheric and Solar-Terrestrial Physics*. s 105–106. 199–206. <https://doi.org/10.1016/j.jastp.2013.10.003>.
- Valladares, C. E. and Chau, J. L. (2012). The low-latitude ionosphere sensor network: Initial results. *Radio Science*, 47(04), 1-18.
- Van Dierendonck, A. J., Klobuchar, J. and Hua, Q. (1993). Ionospheric scintillation monitoring using commercial single frequency C/A code receivers. Paper presented at Proceedings of ION GPS-93, Salt Lake City, UT., 1333-1342.
- Vani, B. C., Shimabukuro, M. H., and Monico, J. F. G. (2017). Visual exploration and analysis of ionospheric scintillation monitoring data: the ISMR query tool. *Computers & Geosciences*, 104, 125-134.

SERVIÇO WEB PARA EXTRAÇÃO DE MÉTRICAS FENOLÓGICAS PARA AGRICULTURA USANDO SÉRIES TEMPORAIS DE IMAGENS DE SATÉLITE

Gabriel K. Alves^{1,2}, Karine R. Ferreira¹, Bruno Schultz², Luiz F. R. Carvalho²,
Thiago G. Rodrigues²

¹Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758 – 12227-010 – São José dos Campos – SP – Brasil

²Visiona Tecnologia Espacial, São José dos Campos, SP, Brasil

{gabriel.alves, karine.ferreira}@inpe.br

{gabriel.alves, bruno.schultz, thiago.rodrigues}@visionaespecial.com.br

lf.aerospace@gmail.com

Resumo. O estudo dos padrões de vegetação sazonal observados por sensoriamento remoto é chamado de *Land Surface Phenology (LSP)*. A partir de imagens de sensoriamento remoto, é possível obter métricas usadas para o monitoramento fenológico que auxiliam no entendimento da dinâmica da vegetação e na tomada de decisão. Existem diferentes métodos presentes na literatura para a extração dessas métricas a partir de imagens de satélites. Porém, um dos desafios é a extração dessas métricas a partir dos grandes volumes de imagens disponibilizadas atualmente por diferentes provedores. Especialistas se deparam com limitações de hardware para processar esse grande volume de dados em computadores pessoais. Para isso, o objetivo deste trabalho é desenvolver um serviço web para extração de métricas fenológicas a partir de grandes volumes de imagens modeladas como cubos de dados multidimensionais e de séries temporais de índices de vegetação. Esse serviço segue uma arquitetura “cliente-servidor”, processando todo o dado do lado do servidor e retornando para o cliente apenas o resultado do processamento. Usando esse serviço, um especialista pode extrair métricas a partir de grandes volumes de imagens sem se preocupar com limitações de processamento e com instalações de pacotes e sistemas em seu computador pessoal.

1. Introdução

Fenologia é o estudo da sazonalidade dos eventos cíclicos de vida das espécies. Na agricultura, o conhecimento sobre a sazonalidade da vegetação é muito útil para definir, por exemplo, o uso de fertilizantes e de irrigação [Zeng et al. 2020].

A fenologia da vegetação medida a partir de imagens de sensoriamento remoto é chamada de *Land Surface Phenology (LSP)* [Noormets 2009]. A maioria das técnicas de LSP extraem métricas relacionadas ao crescimento e desenvolvimento da vegetação a partir de séries temporais de índice de vegetação. Alguns exemplos dessas métricas são: o início da estação (SOS), o pico da estação de crescimento (PEAK), o final da

estação (EOS) e a duração da estação (LOS). Na agricultura, estas métricas estão associadas à emergência foliar, ao máximo desenvolvimento vegetativo, à colheita das culturas agrícolas e à duração do ciclo da cultura, respectivamente [Zeng et al. 2020].

Existem diferentes métodos presentes na literatura para a extração dessas métricas fenológicas a partir de imagens de satélites e séries temporais de índice de vegetação, como o *Normalized Difference Vegetation Index* (NDVI). Esses métodos incluem desde abordagens simples que utilizam janelas de suavização linear até métodos mais complexos, como ajuste de funções de curva analíticas [Zeng et al. 2020].

O TIMESAT e o SEN2AGRI [Jönsson and Eklundh 2004, Sen2Agri] são exemplos de sistemas usados para extração de métricas fenológicas a partir de imagens de satélites. *CropPhenology*, *Phenofit* e *GreenBrown* [Araya et al. 2018, Zhang et al. 2018, Forkel et al. 2013] são exemplos de pacotes na linguagem R que produzem métricas fenológicas para agricultura a partir de séries temporais NDVI. O pacote *CropPhenology* extrai 15 métricas baseadas em fenologia tanto de séries temporais de imagens quanto de um conjunto de imagens multi-temporais no formato *raster* [Araya et al. 2018].

Recentemente, satélites de observação da Terra têm produzido grandes volumes de imagens com diferentes resoluções espaciais e temporais. O programa Copernicus da Comissão Europeia produz aproximadamente 10 TB de dados do satélite Sentinel por dia [Kempeneers and Soille 2017]. Portanto, um dos desafios encontrados atualmente é a extração de métricas fenológicas a partir desse grande volume de imagens. Especialistas se deparam com limitações de hardware para processar esse grande volume de dados em computadores pessoais.

O objetivo deste trabalho é projetar e desenvolver um serviço web para extração de métricas fenológicas a partir de grandes volumes de imagens modeladas como cubos de dados de observação da Terra e de séries temporais de índices de vegetação, com foco em aplicações agrícolas. Esse serviço segue uma arquitetura “cliente-servidor” que permite que os dados sejam processados do lado do servidor sem que o cliente precise baixar estes dados em sua máquina pessoal. Usando esse serviço, um especialista pode extrair métricas a partir de grandes volumes de imagens sem se preocupar com limitações de processamento e com instalações de pacotes e sistemas em seu computador pessoal.

O desenvolvimento de componentes de software como serviços web tem várias vantagens, como facilidade de uso e interoperabilidade entre sistemas. Serviços web facilitam a integração entre diferentes linguagens de programação, como R e Python, e diferentes tipos de aplicativos clientes [Vinhas et al. 2017]. Além disso, esses serviços podem disponibilizar recursos com um alto nível de abstração para os usuários, que não precisam se preocupar com computadores de processamento, instalação de softwares ou ambientes de programação.

Nesse contexto, esse artigo descreve o serviço web proposto, sua arquitetura, sua implementação utilizando os dados do projeto Brazil Data Cube (BDC) do INPE e os resultados preliminares alcançados até o momento.

2. Serviço web para extração de métricas fenológicas

A Figura 1 mostra a arquitetura do serviço web para extração das métricas fenológicas. O acesso e recuperação das séries temporais NDVI são feitos através dos serviços WTSS

(*Web Time Series Service*) e STAC (*Spatiotemporal Asset Catalog*) do projeto BDC [Ferreira et al. 2020].



Figura 1. Arquitetura do serviço web para extração de métricas fenológicas.

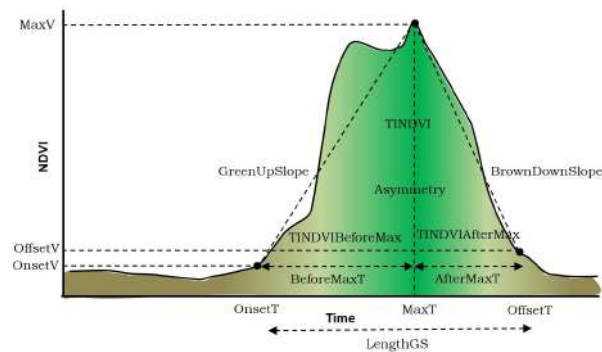
A etapa de pré-processamento consiste em uma normalização dos dados NDVI *int16* com valores $min=-10000.0$ e $max=10000.0$ para *float64* com valores $min=-1.0$ e $max=1.0$. Logo após, é feito um mascaramento de nuvens através de uma banda de qualidade fornecida pelos serviços BDC produzidas por diferentes métodos (*Fmask 4*, *MAJA*, *Sen2Cor* e *s2cloudless*) para cada tipo de sensor [Sanchez et al. 2020].

Após mascarar as nuvens e sombras de nuvens das séries temporais, é realizado uma interpolação linear simples para redução de ruídos referentes a ausência de dados observados. E por fim, a última etapa do pré-processamento é calcular a mediana das séries temporais de uma região específica (por exemplo, uma região de um talhão agrícola) para obter uma melhor representação espacial do conjunto de dados.

Para a extração das métricas fenológicas, foram adotados os métodos do pacote *CropPhenology* em R, e *Change Detection Derivated* que usam cálculos baseados na derivada da curva espectral para detectar os pontos de máximo e mínimo [Araya et al. 2018, Zeng et al. 2020]. O *CropPhenology* extrai 15 métricas fenológicas: *OnsetV*, *OnsetT*, *MaxV*, *MaxT*, *OffsetV*, *OffsetT*, *LengthGS*, *BeforeMaxT*, *AfterMaxT*, *GreenUpSlope*, *BrownDownSlope*, *TINDVIBeforeMax*, *TINDVIAfterMax*, *TINDVI* e *Asymmetry* [Araya et al. 2018]. Figura 2 mostra a representação das métricas na curva NDVI.

O serviço web proposto trabalha com padrões de comunicação REST usando a linguagem universal de troca de dados com o formato JSON, *JavaScript Object Notation* e propõe cinco operações, que podem ser vistos na Figura 3:

1. *pheno-metrics*: Esta operação é derivada do pacote *CropPhenology* e recebe como parâmetro: área de interesse, período de análise, token de acesso e nome do cubo de dados fornecidos pelo BDC. E retorna além das 15 métricas propostas pelo pacote em formatos raster *.tiff* e *.grd*, as informações sobre os dados analisados.
2. *single-phenology*: Também é uma função derivada do *CropPhenology* que recebe os mesmos parâmetros, porém calcula as métricas através de um único ponto em



Fonte: [Araya et al. 2018].

Figura 2. Curva NDVI com representação das métricas fenológicas obtidas pelo *CropPhenology*.

uma única estação. Retorna 15 métricas propostas pelo pacote, as informações sobre a série usada e recursos do BDC. Um exemplo de sua resposta pode ser vista na Figura 4.

3. *pheno-metrics-derived*: Função implementada através do *Change Detection Derived*, que leva em conta os princípios de derivada e pontos locais de máximo e mínimo. Recebe os mesmos parâmetros que as demais operações e retorna 3 métricas fenológicas.
4. *get-thumbnail*: Operação auxiliar com a finalidade de retornar os thumbnails dos rasters gerados pelo *endpoint pheno-metrics* no formato *image/png*.
5. *get-image*: Operação auxiliar com a finalidade de retornar os rasters gerados pelo *endpoint pheno-metrics* no formato *image/tiff* ou *grd-file*.

Extract Phenologic metrics

1.0.0 OAS3

<http://127.0.0.1:8080/openapi.json>

API for extract time series

Servers

Metrics

Operations for extract phenologic metrics

GET **/pheno-metrics** Extract 15 phenologic metrics from multi-temporal IV raster stack.

GET **/single-phenology** Extract 15 phenologic metrics from VI time series for a single season

GET **/pheno-metrics-derived** Extract 3 phenologic metrics from VI time series using change detection derived method.

Aux Functions

Auxiliary functions

GET **/get-thumbnail** Returns the thumbnail (.png) of the /pheno-metrics endpoint.

GET **/get-image** Returns tile (.tiff/.grd/.grt) of endpoint /pheno-metrics.

Figura 3. Documentação das operações disponíveis no serviço web.

```
SinglePhenology {
  area: string
  example: {"type": "Polygon", "coordinates": [[[-42.0968627929688, -19.6478313116844], [-42.0968627929688, -19.6478313116844]]]}
  datacube: string
  example: S2-SEN2COR_10_16D_STK-1
  period: string
  example: 2019-01-01/2020-01-01
  source: string
  example: https://brazildatacube.dpi.inpe.br/utss
  service: string
  example: Rutss
  version: string
  example: 0.9.2
  method: string
  example: single-phenology
  time_series: > {...}
  phenologic_metrics: > {...}
  statistics: > {...}
}
```

Figura 4. Esquema resposta da operação *single-phenology*.

3. Estudo de caso e resultados preliminares

Para testar o serviço proposto, foi realizado um estudo de caso utilizando talhões agrícolas no estado do Maranhão e selecionados 10 pontos aleatórios inseridos em cada um deles para a extração das séries temporais. O cubo de dados usado foi o S2-SEN2COR_10_16D_STK1 para um período de 2019/2020. A Figura 5 mostra os talhões usados juntamente com a mediana das séries temporais extraídas de cada talhão.

A função *single-metrics-derivated* do serviço que usa o método *Change Detection Derivated* foi utilizada para detectar as métricas SOS, PEAK e EOS a partir da mediana das séries obtidas, a técnica foi capaz de extrair um estágio de cultivo retornando os pontos SOS, PEAK e EOS. Os resultados podem ser observados na Figura 6.

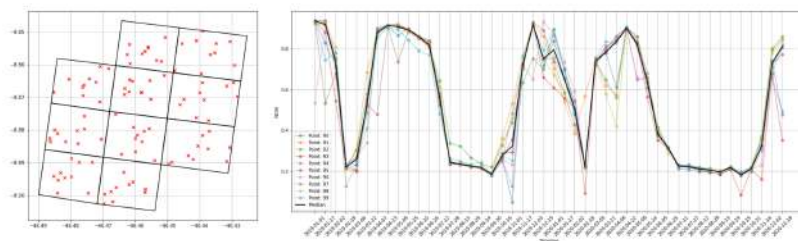


Figura 5. Talhões usados e mediana das séries temporais extraídas.

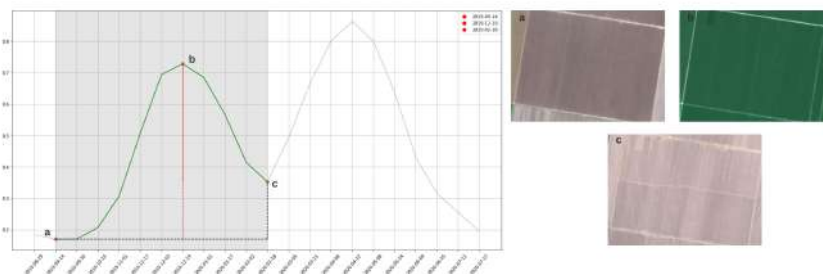


Figura 6. Métricas fenológicas SOS (a), PEAK (b) e EOS (c) detectadas pela função *pheno-metrics-derivated* do serviço que utiliza o método *Change Detection Derivated* e suas respectivas imagens a direta.

Para o desenvolvimento do serviço web, foram usados as linguagens de programação *Python* e *R*, tecnologias de empacotamento de código de softwares com o *Docker*. Pacotes para geração de API (*Plumber* e *Flask*), além de seguir padrões e especificações de documentação como o *Openapi* e *Swagger*.

Este é um trabalho em desenvolvimento e, portanto, o serviço não está em operação ainda. Os próximos passos desse trabalho são: (1) explorar meios de implantação do serviço e avaliar seu desempenho; (2) testar novos métodos e pacotes em *R* como *GreenBrown* e *Phenofit*; (3) usar a nova versão do serviço WTSS para extração de séries temporais de polígonos para que não haja necessidade de criar pontos aleatórios nas áreas de interesse e (4) aprimorar o serviço para melhorar a sua performance.

Referências

- Araya, S., Ostendorf, B., Lyle, G., and Lewis, M. (2018). Crop phenology: An r package for extracting crop phenology from time series remotely sensed vegetation index imagery. *Ecological Informatics*, 46:45–56.
- Ferreira, K. R., Queiroz, G. R., Vinhas, L., Marujo, R. F. B., Simoes, R. E. O., Picoli, M. C. A., Camara, G., Cartaxo, R., and et al (2020). Earth observation data cubes for brazil: Requirements, methodology and products. *Remote Sensing*, 12(24).
- Forkel, M., Carvalhais, N., Verbesselt, J., Mahecha, M. D., Neigh, C. S., and Reichstein, M. (2013). Trend change detection in ndvi time series: Effects of inter-annual variability and methodology. *Remote Sensing*, 5(5):2113–2144.
- Jönsson, P. and Eklundh, L. (2004). Timesat—a program for analyzing time-series of satellite sensor data. *Computers & geosciences*, 30(8):833–845.
- Kempeneers, P. and Soille, P. (2017). Optimizing sentinel-2 image selection in a big data context. *Big Earth Data*, 1(1-2):145–158.
- Noormets, A. (2009). *Phenology of ecosystem processes: applications in global change research*. Springer.
- Sanchez, A. H., Picoli, M. C. A., Camara, G., Andrade, P. R., Chaves, M. E. D., Lechler, S., Soares, A. R., Marujo, R. F. B., Simões, R. E. O., Ferreira, K. R., and Queiroz, G. R. (2020). Comparison of cloud cover detection algorithms on sentinel–2 images of the amazon tropical forest. *Remote Sensing*, 12(8).
- Sen2Agri. Sentinel-2 for agriculture. <http://www.esa-sen2agri.org/>. Accessed: 2022-09-02.
- Vinhas, L., de Queiroz, G. R., Ferreira, K. R., and Camara, G. (2017). Web services for big earth observation data. *Revista Brasileira de Cartografia*, 69(5).
- Zeng, L., Wardlow, B. D., Xiang, D., Hu, S., and Li, D. (2020). A review of vegetation phenological metrics extraction using time-series, multispectral satellite data. *Remote Sensing of Environment*, 237:111511.
- Zhang, Q., Kong, D., Shi, P., Singh, V. P., and Sun, P. (2018). Vegetation phenology on the qinghai-tibetan plateau and its response to climate change (1982–2013). *Agricultural and Forest Meteorology*, 248:408–417.

Modeling and Validation of Elevation Data Obtained by GNSS and Drone Devices for Managing Floodings in Coastal Lands Caused by Climate Change

Carlos A. Felgueiras¹, Claudia C. Santos¹, Laércio M. Namikawa¹, Paulo C. G. Albuquerque¹, Sérgio Rosim¹, Sidnei J. S. Santanna¹

¹Instituto Nacional de Pesquisas Espaciais (INPE) – São José dos Campos – SP

{carlos.felgueiras,claudia.santos,laercio.namikawa,sergio.rosim,sidnei.santanna}@inpe

Abstract. *This short-paper reports and explores a methodology to model and validate geographical elevation information from samples, points and images, acquired by Global Network Satellite System (GNSS) and Aerial Drone devices. The elevation models will be used on flooding management activities of coastal lands caused by climate change. The methodology of this article addresses details: on how to obtain and process GNSS sample points and drone images to generate the final Digital Elevation Model (DEM) and; on how to validate the DEM information to attain its quantitative quality. Also, the article presents some initial ideas for creating different flooding scenarios according to the sea level variation. The simulated scenarios can be used on decision making activities to minimize risks of human and material losses. The up-to-now results, obtained in a case study in **Guaecá** beach area, show the applicability and the feasibility of the proposed methodology.*

1. Introduction

Climate change has pointed, according to the consensus of many serious researchers, to an increase in the planet's temperature [Hansen et al. 2006, Menne et al. 2018]. One of the effects mentioned is the possible rise in sea level that can cause, among other problems, floods in coastal areas, notably in urban areas [Brown et al. 2020, Frederikse et al. 2020, Hinkel et al. 2019] [Brow et al. 2020, Frederikse et al. 2020, Hinkel et al. 2019]. The delimitation of the area to be flooded is crucial for scenario simulations that indicate the loss of urban area with large population density and, usually, with high economic value. In Brazil, 60% of the population lives in the coastal range [Marengo and Scarabi 2016]. Significant effects also occur in non-urban coastal areas, such as those with coconut plantations, for example.

The scenarios' creation simulating increases in the current level of the oceans is essential for public managers to make preventive decisions to avoid these floods, or even mitigate their effects, as the economic, social and environmental damages can be quantified with the scenarios. The work of delimiting these flooded areas begins with the creation of a Digital Elevation Model (DEM), a relief representation, of the geographic region of interest. Thus, the DEM to be used to accomplish reliable simulation processes must be of the best possible cartographic quality considering the spatial position and, mainly, the elevation information. Shuttle Radar Topography Mission (SRTM) [Rodriguez et al. 2005] and ASTER Global Digital Elevation Model (GDEM)

[Abrams et al. 2020] are examples of DEMs that can be found for free in the internet. Unfortunately, for high quality simulations, those data are not suitable. DEMs of high spatial and altimetric accuracy can be created from a set of sample points obtained using GNSS instruments and/or a set of drone images [Sze et al. 2015, Bolkas et al. 2016].

In this context, the objective of this article is to report and to explore a methodology to model and validate elevation data addressing the topics of how to obtain and process GNSS sample points and drone images to generate a high accurate DEM and of how to validate the DEM information to assess its quantitative quality. Geostatistical and interferometry procedures are used to create the DEMs while direct and cross validations are considered to evaluate their global accuracies. A case study in the beach of Guaecá, located in the city of São Sebastião, on the northern coast of the State of São Paulo, Brazil, is used to illustrate the applicability of the proposed methodology. Also, results of the case study along with related discussion are presented.

2. Material and Methods

The input data in this research are a sample set of elevation points and a set of high-resolution images acquired in a geographical region of interest. The sample set was collected using Global Network Satellite System (GNSS) instruments while the images were acquired by drone flights carried during field works. The sample points were determined on-site through three GNSS geodetic equipment: Sokkia Stratus (L1 frequency), Topcom Hiper (L1/L2 frequency) and Trimble R-4 (L1/L2 frequency). The images were gathered on aerophotogrammetric surveys and conducted using a quadcopter multimotor UAV model Phantom 3 Advanced drone, equipped with a Sony Exmor sensor with 12.76 megapixels and focal distance of 3.61 mm.

The methodology of this research has the following steps: acquisition of a set of elevation points using GNSS equipment; gathering local images by drone flights; processing the GNSS raw data to obtain orthometric referenced information, hereafter named GNSS samples; processing the drone images, with the support of the GNSS samples, to obtain the drone models of the mapped region, the Surface DEM (SDEM) and the Terrain DEM (TDEM); creation of a DEM model using the GNSS samples, hereafter named GNSS DEM, and geostatistical tools known as variography analysis and kriging prediction and; validations of the GNSS and the Drone DEMs by means of direct and cross validation procedures. The methodology above was applied in a case study and the results are presented and analyzed in the next sections of this article.

The geometric heights of the raw GNSS data, were transformed into geoidal orthometric heights, based on the reference geocentric ellipsoid GRS80, in the software MAPGEO 2015 [Blitzkow et al. 2016]. The Drone DEMs were created using the PIX4D [Barbasiewicz et al. 2018] software. The PyESSDAES software [Felgueiras et al. 2019] was used to generate and validate de GNSS DEM. The Geographical Information System (GIS) SPRING [Câmara et al. 1996] was used as the spatial database for visualizations and complementary analyses of the GNSS sample set and the resulting DEM maps.

3. A Case Study

The proposed methodology was applied in the coastal region known as Guaecá beach located in the municipality of São Sebastião, on the northern coast of São Paulo state,

Brazil. Figure 1 presents the Guaecá region, with latitudes from 23° 49' 00" S to 23° 49' 27" S and longitudes from 45° 27' 50" W to 45° 27' 09" W.



Figure 1. Guaecá region (Source: Google Maps)

Figure 2 shows the spatial distribution of points of the GNSS sample set, marked as plus signals, along with their elevation values, plotted in the SPRING software. Elevation statistical values of the 57 sample points are: minimum 1.03m, maximum 6.15m, mean 3.95, variance 1.70m, standard deviation 1.30m, coefficient of variation 0.33m, skewness -0.66m and kurtosis -0.15m. All the sample points coordinates, horizontal, vertical and altimetric, present accuracy better than 10cm.

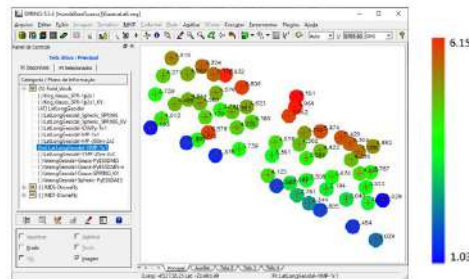


Figure 2. GNSS sample set of the Guaecá region

Figure 3 illustrates the drone image acquisition planning showing, in yellow lines, the programmed flight regions.



Figure 3. Drone flight areas in the Guaecá region

4. Results and Discussion

This section presents some results and discussion related to the modelling and validation of the DEMs created for the case study region considered in this article.

4.1. Digital Elevation Model from the GNSS sample set

Figure 4, shows results of the semivariogram analysis of the GNSS sample set and the predicted map, the GNSS DEM of the Guaecá region, obtained in the PyESSDAES. The theoretical semivariogram of Figure 4(a) was fitted with a Gaussian function having 0.12m of nugget effect, 1.55m of contribution and 296.5m of range values. The map of Figure 4(b) was accomplished with the geostatistical estimator, known as ordinary kriging approach, creating a grid of 798 columns x 746 rows with a spatial resolution of 1m x 1m.

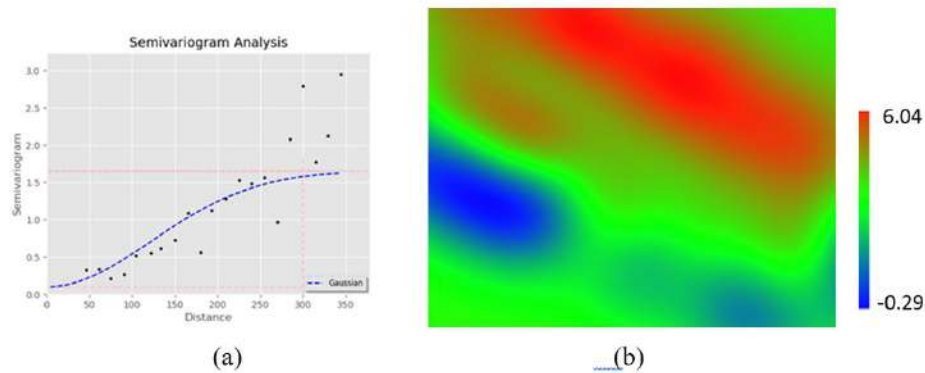


Figure 4. The semivariogram of the GNSS sample st and (b) the DEM of Guaecá region

Visually, the general aspect of kriging estimated map agrees with the circle colors of the sample set presented in Figure 2. Also, the estimated map, as expected, presents a smooth surface representing the tendency of the elevation sample set information.

4.2. Digital Elevation Model from Drone Images

The resulting maps of the processed drone images are presented in Figure 5. The map of Figure 5(a) is the Surface model, hereafter named as Drone SDEM, while the map of Figure 5(b) is the Terrain model, hereafter named Drone TDEM. The SDEM contains extra terrain information, as edifications and vegetations, that were filtered, inside the Pix4D the software, to generate the TDEM information.

Although the Figure 5(a) presents more details, more red spots for example, it cannot be suitable to applications that require only the terrain information.

4.3. Direct and Cross Validations of the DEMs

The GNSS DEM and the Drone DEMs were validate using cross and direct validations to assess the quantitative quality of the generated elevation Models.

The cross validation was performed only for the kriged GNSS DEM. In this process, elevation values are predicted from GNSS DEMs not containing the sample in consideration. Differences among actual elevation and respective predict sample values are

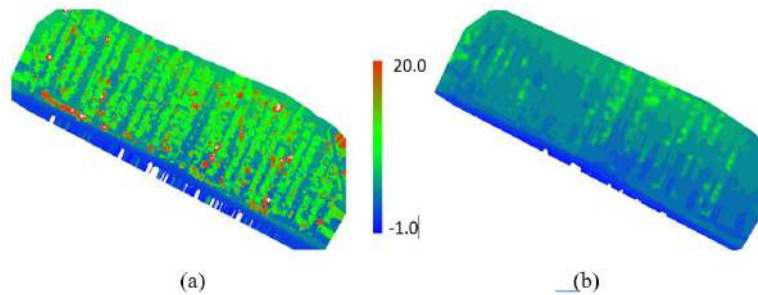


Figure 5. Drone Models of Guaecá region: (a) Surface Model (Drone SDEM) and (b) Terrain Model (Drone TDEM)

considered to obtain statistics and Root Mean Squared (RMS) error values of the differences. The GNSS sample set as the test data for the direct validations was used in this research. Although this is biased, because the same sample set was used to obtain the DEMs, this was the only source data available up to now for this task. In the near future a specific field work will be carried out to get independent test data for the validations.

Table 1 reports the results of the DEMs' validations using cross and direct validations for the kriged GNSS DEM and direct validations for the drone DEMs.

Table 1. DEM Validation results

		Statistics / Errors (m)			
		Mean	Variance	Std. Deviation	RMS
DEM	GNSS DEM Cross	-0.007	0.77	0.88	0.88
	GNSS DEM	0.19	0.34	0.58	0.61
	Drone SDEM	0.04	0.24	0.49	0.49
	Drone TDEM	-0.40	0.35	0.59	0.71

It can be seen by analyzing the results presented in Table 1 that quantitatively the Drone SDEM is the best model obtained in this study. This can be explained by the fact that it uses the Drone data besides the GNSS samples. Moreover the GNSS DEM represents only the tendency elevation and the Drone TDEM was derived from the application of a filter in the Drone SDEM data. Better results are expected when using the independent set of test samples that will have accuracy error lower than 10cm.

5. Final Remarks

This short paper presents and explores, with a case study, a methodology to model and validate elevation information from samples obtained with GNSS and Drone devices. The outcomes demonstrate the applicability and the feasibility of the proposed methodology. Also, the quantitative results of the DEMs' validations show that the use of drone information, along with GNSS basic data, increases the DEM accuracy.

This research is the first step of a more complex experiment that aims to use DEMs as basic input data for flooding management activities of coastal lands caused by climate

change. The idea is to propose a complete methodology that allows to delimit flooding areas in coastal regions, mainly the urban ones, considering the sea level rising.

References

- Abrams, M., Crippen, R., and Fujisada, H. (2020). Aster global digital elevation model (gdem) and aster global water body dataset (astwbd). *Remote Sensing*, 12(7).
- Barbasiewicz, A., Widerski, T., and Daliga, K. (2018). The analysis of the accuracy of spatial models using photogrammetric software: Agisoft photoscan and pix4d. *E3S Web of Conferences*, 26:00012.
- Blitzkow, D., Matos, A. C. O. C. d., Machado, W. C., Nunes, M. A., Lengruber, N. V., Xavier, E. M. L., and Fortes, L. P. S. (2016). Mapgeo2015: O novo modelo de ondulação geoidal do brasil. *Revista Brasileira de Cartografia*, 68(10).
- Bolkas, D., Fotopoulos, G., Braun, A., and Tziavos, I. N. (2016). Assessing digital elevation model uncertainty using gps survey data. *J. of Surveying Engineering*, 142(3).
- Brown, S., Wadey, M., Nicholls, R., Shareef, A., Khaleel, Z., Hinkel, J., Lincke, D., and McCabe, M. V. (2020). Land raising as a solution to sea-level rise: an analysis of coastal flooding on an artificial island in the maldives. *Journal of Flood Risk Management*, 13(s1).
- Felgueiras, C. A., de Oliveira Ortiz, J., and Camargo, E. G. (2019). An user-friendly python application for exploratory and structural spatial dependence analysis for sample points of spatial attributes. In Filho, J. L. and Monteiro, A. M. V., editors, *XX Brazilian Symposium on Geoinformatics - GeoInfo 2019, São José dos Campos, SP, Brazil, November 11-13, 2019*, pages 25–35. MCTIC/INPE.
- Frederikse, T., Landerer, F., Caron, L., Adhikari, S., Parkes, D., Humphrey, V. W., Danggendorf, S., Hogarth, P., Zanna, L., Cheng, L., et al. (2020). The causes of sea-level rise since 1900. *Nature*, 584(7821):393–397.
- Hansen, J., Sato, M., Ruedy, R., Lo, K., Lea, D. W., and Medina-Elizade, M. (2006). Global temperature change. *Proc. Natl. Acad. Sci.*, 103:14288–14293.
- Hinkel, J., Church, J. A., Gregory, J. M., Lambert, E., Le Cozannet, G., Lowe, J., Mcinnes, K., Nicholls, R., van Der Pol, T. D., and van de Wal, R. S. W. (2019). Meeting User Needs for Sea Level Rise Information: A Decision Analysis Perspective. *Earth's Future*, 7(3):320–337.
- Marengo, J. A. and Scarabi, F. R. (2016). Impacto, vulnerabilidade e adaptação das cidades costeiras brasileiras às mudanças climáticas. Technical report, COPPE - UFRJ. Rio de Janeiro, Brasil.
- Menne, M. J., Williams, C. N., Gleason, B. E., Rennie, J. J., and Lawrimore, J. H. (2018). The global historical climatology network monthly temperature dataset, version 4. *Journal of Climate*, 31(24):9835 – 9854.
- Sze, L. T., Cheaw, W. G., Ahmad, Z. A., Ling, C. A., Chet, K. V., Lateh, H., and Bayuaji, L. (2015). High resolution dem generation using small drone for interferometry sar. In *2015 International Conference on Space Science and Communication (IconSpace)*, pages 366–369. IEEE.

Pacote ‘Contraste’: aplicação de técnicas de contraste baseado no uso e cobertura da terra

**Diego M. de Melo^{1,2}, Gisele Milare^{1,3}, Marielcio G. Lacerda¹,
Gilberto Ribeiro de Queiroz¹, Thales Sehn Körting¹**

¹Instituto Nacional de Pesquisas Espaciais (INPE)
Av. dos Astronautas, 1758. Jd. da Granja – 12227-010 – S. J. Campos - SP - Brasil

²Fundação Universidade Federal do ABC (UFABC)
Av. dos Estados, 5001 - Bairro Bangu – 09210-580 – Santo André - SP - Brasil

³Instituto Nacional de Colonização e Reforma Agrária (INCRA)
302 Norte Alameda 01 – 77006-3636 – Palmas - TO - Brasil

diego.marcoccio@ufabc.edu.br, {gisele.milare, marielcio.lacerda, gilberto.queiroz, thales.korting}@inpe.br

Abstract. *Image contrast is a useful tool for visual analysis in remote sensing. Seeking improvements in the use of this tool, in this paper a package has been proposed, in python language, to automatically apply the contrast according to the land use and land cover class of each image region, independently. For this, land use and land cover data were used from the Mapbiomas Project, OLI/Landsat 8 images and selected six contrast techniques. Each technique was applied on image containing at least 5 land use and land cover classes and selected the best contrast technique for each class. The custom contrast function has been implemented selection according to the user's choice.*

Resumo. *O contraste de imagens é uma ferramenta útil para a realização de análises visuais em sensoriamento remoto. Buscando melhorias na utilização dessa ferramenta, neste trabalho foi proposto um pacote, em linguagem python, para aplicação automática do contraste conforme a classe de uso e cobertura da terra de cada região da imagem, de forma independente. Para isso, foram utilizados dados de uso e cobertura da terra do Projeto Mapbiomas e imagens do sensor OLI/Landsat 8. Seis técnicas de contraste foram selecionadas e aplicadas em imagem contendo as 5 classes de uso e cobertura da terra, sendo selecionada a melhor técnica de contraste para cada classe. Também implementou-se a função para escolha personalizada de acordo com a opção do analista.*

1. Introdução

Imagens de satélite proporcionam uma visão multi-temporal de extensas áreas da superfície terrestre, revelando as transformações do ambiente, além dos impactos causados por fenômenos naturais e atividades antrópicas, como no uso e ocupação da terra [Florenzano 2007]. A utilização de imagens de satélite para análises e monitoramento do uso e cobertura da terra tem se tornado frequente. Como exemplo, cita-se o aumento expressivo no uso de imagens Landsat, com mais de 100.000 imagens utilizadas em estudos em 2020 [Hemati et al. 2021].

Para interpretação dessas imagens, o analista se utiliza da análise das propriedades dos elementos da imagem, como cor e tonalidade, localização, tamanho, forma e textura [Jensen 2009], determinando assim quais serão as suas classes. Entretanto, o contraste entre os alvos pode não ser totalmente percebido pelo analista, pois as variações do nível de cinza dos pixels das imagens são restritas a uma pequena diferença digital [Meneses and Almeida 2012].

Para melhorar a visualização, algumas técnicas de contrastes são aplicadas à imagem, através de operações matemáticas sobre os valores digitais dos pixels. Dessa forma, algoritmos que aplicam melhoria na visualização das imagens sem alteração de seu conteúdo são relevantes para aplicações em sensoriamento remoto. Todavia, há casos em que a imagem de sensoriamento remoto possui diversidade de alvos e, eventualmente, uma única técnica de contraste aplicada em toda a imagem pode ressaltar mais um alvo, mascarando outros. A aplicação de contraste por classe de uso e cobertura do solo pode ser uma saída para ressaltar todos os alvos presentes na imagem.

Nesse contexto, esse trabalho tem os seguintes objetivos: (1) selecionar a técnica de contraste, aplicada às imagens do sensor *Operational Land Imager* (OLI) do satélite Landsat 8, visualmente adequada para cinco classes selecionadas de uso e cobertura da terra do projeto Mapbiomas; (2) implementar um pacote, em linguagem *python*, com módulos para cada técnica de contraste, aplicando essas técnicas por classe, de duas formas: automática e personalizada.

2. Material e métodos

2.1. Área de estudo e dados utilizados

A área de estudo é uma região localizada no município de São Félix do Xingu, Estado do Pará, Brasil (Figura 1).

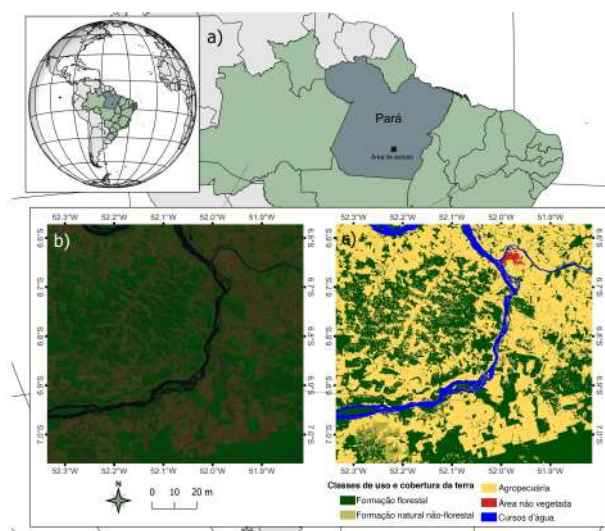


Figura 1. a) Área de estudo no Estado do Pará, Brasil; b) Imagem Landsat 8 R(6) G(5) B(4); c) Classes de uso e cobertura da terra [Mapbiomas 2020].

Foram utilizados dados de uso e cobertura da terra do Projeto Mapbiomas [Mapbiomas 2020], selecionando cinco classes: formação florestal, formação natural não florestal, agropecuária, área não vegetada e corpos d'água. Utilizou-se as bandas SWIR (6), NIR (5) e RED (4), na composição de cores falsas em RGB, respectivamente, da imagem do sensor OLI, a bordo do Landsat 8 (órbita/ponto: 225/065 com data de aquisição: 23/07/2018), para aplicar técnicas de contraste.

2.2. Técnicas de contraste

As técnicas de contraste, normalmente empregadas como etapa de pré-processamento no processamento de imagens, visam melhorar a qualidade das imagens, a fim de aumentar a discriminação visual entre os objetos contidos na imagem, facilitando a percepção de detalhes para o olho humano. No geral, a escolha de um realce de contraste apropriado é essencialmente empírico. Dentre seis técnicas de contraste (Tabela 1), através de análise visual, selecionamos a técnica que melhor se destacou para cada uma das cinco classes de uso e cobertura da terra.

Tabela 1. Técnicas de contraste selecionadas para aplicação em classes de uso e ocupação da terra.

Técnica de contraste	Descrição
Raiz quadrada	$y = A * \sqrt{x}$, onde: y = novo valor de nível de cinza; x = valor original de nível de cinza; and A = Fator de ajuste.
Função quadrática	$y = A * x^2$, onde: y = novo valor de nível de cinza; x = valor original de nível de cinza; and A = Fator de ajuste.
Linear	$y = Ax + B$, onde: y = novo valor de nível de cinza; x = valor original de nível de cinza; A = Inclinação da linha (ângulo tangente); and B = Fator de incremento, limites mínimos e máximos definidos pelo analista
Mínimo/Máximo	$y = Ax + B$, onde: y = novo valor de nível de cinza; x = valor original de nível de cinza; A = Inclinação da linha (ângulo tangente); and B = Fator de incremento, limites mínimos e máximos definidos pelo analista
Negativo	$y = -(Ax + B)$, onde: Y = novo valor de nível de cinza; x = valor original de nível de cinza; A = Inclinação da linha (ângulo tangente); and B = Fator de incremento, limites mínimos e máximos definidos pelo analista
Equalização de histograma	$y = (f.255)/P$, onde: y = novo valor de nível de cinza; f = frequência acumulativa para o nível de cinza x; and P = número total de pixel.

2.3. Algoritmo de aplicação de contraste por classe de uso e cobertura da terra

Para a criação do pacote 'contraste', desenvolveu-se um algoritmo que permite aplicar técnicas de contraste de forma automática ou customizada, por classes de uso e cobertura da terra. O algoritmo foi implementado utilizando as bibliotecas *gdal*, *rasterio*, *os*, *numpy* e *matplotlib*, presentes na linguagem *python*.

3. Resultados e discussão

As técnicas de contraste para cada classe (Tabela 2) foram selecionadas através de análise visual.

O pacote 'contraste' permite a aplicação de contrastes de forma automática ou personalizada, por classes de uso e cobertura da terra. Na forma automática, o pacote

Tabela 2. Técnica de contraste selecionada para cada classe de uso e cobertura da terra.

Classe de uso e cobertura da terra	Técnica de contraste
Formação florestal	Linear
Formação natural não-florestal	Mínimo/máximo
Agropecuária	Equalização de histograma
Área não vegetada	Mínimo/máximo
Corpos d'água	Equalização de histograma

aplica as técnicas de contraste consideradas mais adequadas por classes através das nomenclaturas utilizada pelo projeto Mapbiomas [Mapbiomas 2020], conforme Tabela 2. Na parte personalizada, o pacote permite ao analista aplicar as técnicas de contraste de forma individualizada. O pacote gerado encontra-se disponível em [Melo et al. 2021].

A Figura 2 apresenta a imagem original e o resultado da aplicação do contraste único para toda a imagem com a técnica raiz quadrada, já a Figura 3 apresenta o resultado da aplicação do contraste otimizado e da aplicação da opção personalizada, com um exemplo da aplicação do contraste Mínimo/Máximo na classe Formação Florestal, e os respectivos histogramas de ambas as imagens.

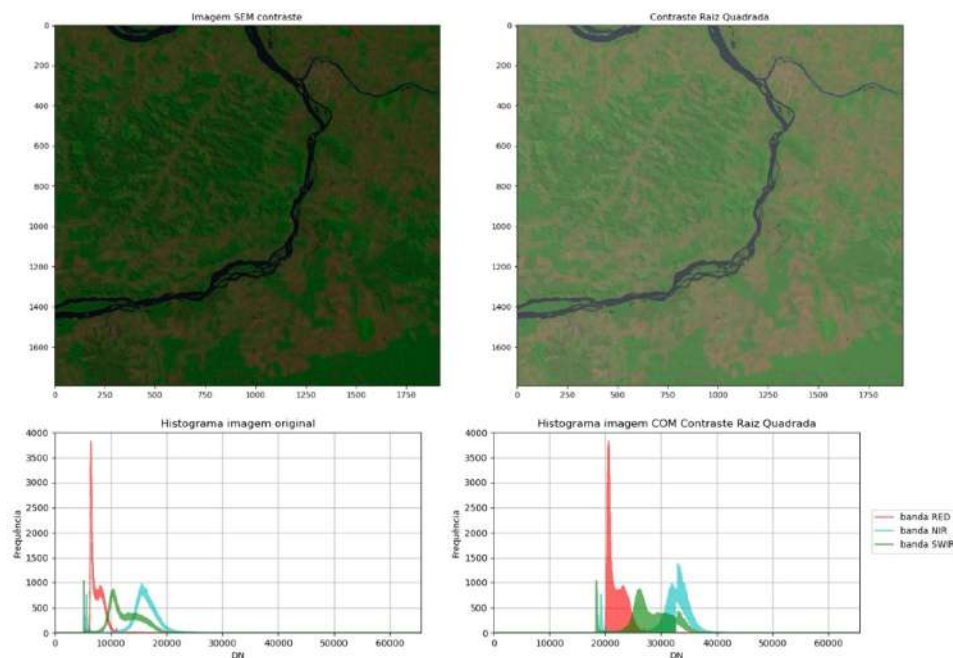


Figura 2. Imagem original, imagem com único contraste aplicado (raiz quadrada) e respectivos histogramas.

A aplicação de técnicas de contraste baseado no uso e cobertura da terra através do pacote 'contraste' possui como limitação a necessidade de um arquivo com as feições de classes de uso e cobertura da terra. Entretanto, essa limitação pode ser contornada com o uso dos dados do Projeto Mapbiomas [Mapbiomas 2020], ou de outra base de dados

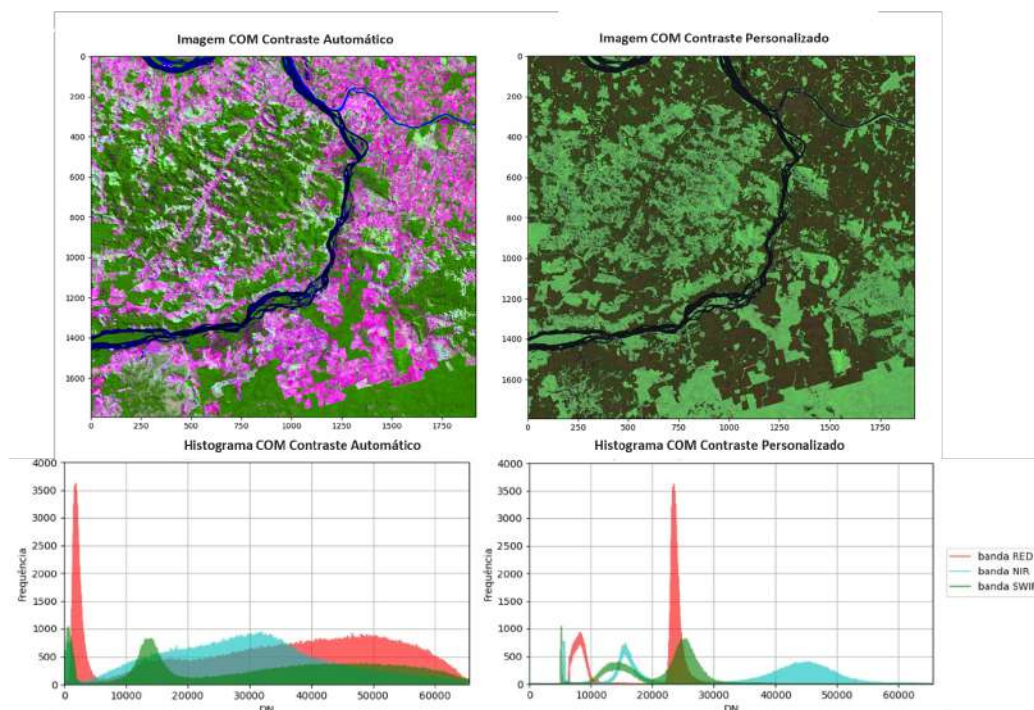


Figura 3. Imagem resultante da aplicação automática de contraste e imagem personalizada gerada pelo pacote 'contraste' e respectivos histogramas.

compatível com as classes definidas no pacote.

4. Considerações Finais

Em sensoriamento remoto, no que se refere à análise visual de imagens de sensores remotos, as técnicas de contraste de imagens são ferramentas úteis para melhorar a discriminação de alvos de interesse. No entanto, nem sempre a aplicação de uma única técnica de contraste gera bons resultados para todas as classes de uso e cobertura da terra nas imagens.

Ao longo deste trabalho, percebe-se que, de acordo com a classe analisada, pode existir uma técnica que oferece um resultado mais adequado em comparação a outras técnicas. Com base na análise visual realizada, conclui-se que para a classe "floresta", a técnica mais indicada foi a de contraste linear, sendo para classe "formação natural não florestal" e "área não vegetada" o contraste mínimo/máximo. Por fim, para as classes "agropecuária" e "corpos d'água", a técnica de equalização do histograma.

O pacote apresentado foi implementado em linguagem *python*, apresentando as formas de aplicação de contraste automática e customizada. A forma automática aplica técnicas de contraste pré-definidas às imagens do satélite Landsat 8, de acordo com as classes de uso e cobertura do Projeto Mapbiomas. No aplicativo personalizado, o analista seleciona a técnica de contraste desejada de forma individualizada para cada classe de uso e cobertura da terra da imagem. Por fim, para trabalhos futuros, sugere-se incluir na aplicação do pacote imagens de outros sensores e outras classes de uso e cobertura da terra não contempladas neste trabalho.

Referências

- Florenzano, T. G. (2007). *Iniciação em sensoriamento remoto*. Oficina de Textos.
- Hemati, M., Hasanlou, M., Mahdianpari, M., and Mohammadimanesh, F. (2021). A systematic review of landsat data for change detection applications: 50 years of monitoring the earth. *Remote sensing*, 13(15):2869.
- Jensen, J. R. (2009). *Sensoriamento remoto do ambiente: uma perspectiva em recursos terrestres*. Parêntese Editora São José dos Campos.
- Mapbiomas (2020). Projeto mapbiomas – coleção 5 da série anual de mapas de cobertura e uso de solo do brasil.
- Melo, D. M., Milare, G., and Lacerda, M. G. (2021). Contraste. <https://github.com/gmilare/Pacote-Contraste->.
- Meneses, P. R. and Almeida, T. d. (2012). Ampliação histográfica de contraste. In Meneses, P. R. and Almeida, T. d., editors, *Introdução ao processamento de imagens de sensoriamento remoto*. Universidade de Brasília, Brasília.

Modeling the dynamics of moving regions for knowledge discovery

Ivanildo Barbosa, Raquel Aparecida Abrahão Costa e Oliveira

¹Seção de Engenharia Cartográfica – Instituto Militar de Engenharia (IME)
Praça General Tibúrcio, 80 – 22.290-270 – Rio de Janeiro – RJ – Brazil

{ivanildo,raquel.oliveira}@ime.eb.br

Abstract. *Several methods to pattern mining in trajectories have been developed in the past years, most of them conceived for moving points. Nevertheless, there are potential demands to analyze the dynamic behavior of features with spatial extents, such as the changes in the land cover and weather imagery. This paper aims to extend both methods and patterns used in trajectory data mining defined for moving points to analyze moving regions. The proposed solution includes the transformation of the set of regions into a set of trajectories, the computation of the geometric differences, and the temporal analysis of both size and complexity of the shapes of such geometries.*

1. Introduction

Objects and phenomena evolve over time, by changing their location, shape, and attributes along their lifetimes. Papers dealing with trajectory mining usually address issues in the movement of point objects (Siabato et al., 2018). Feng and Zhu (2016) provide a comprehensive survey on trajectory data mining based on the definition of trajectory as a sample of location points that the moving object passes. On the other hand, Adrienko et al. (2008) outlines the concept of trajectory as the “path made by the moving entity through the space where it moves”. The geometry is not set in the first moment, however, they cite ahead that “a trajectory can be viewed as a function that matches time moments with positions in space” and “trajectories have to be represented by finite sequences of time-referenced locations”.

McKennney and Frye (2015) define moving regions as “a form of spatiotemporal data in which a region changes in shape and/or position over time”. Earlier, Macedo et al. (2008) adopted the concept of moving regions when the application aims to analyze how the shape of the moving object changes. Junghans and Gertz (2010) highlighted the relevance of the applications involving the mining of moving regions, e.g., wildfires, oil spills, flooding, or clouds of hazardous particles. This list must also consider wide objects detected in video frames as well as the density regions shaped by clustering moving points within a convoy (Jeung et al., 2008). Ortiz, Rocha and Medina (2020) proposed a semantic model for moving regions, aiming to describe the expansion of a region along the main directions (north, south, east and west): they applied such model to describe the expansion of the urban area of the city of Bogotá between 1797 and 1928.

This paper aims to propose an approach to analyze moving regions . It includes a brief analysis of metrics to quantify both the movement of regions and the changes in their shapes. However, it does not include the interactions with other features.

For these reasons, it can be established three premises for the data sets: a) geometries are free of outliers; b) each region and its homologous in the next instant matches correctly; and c) all the regions are homogeneous, i.e., it represents the same intensity of phenomena all over its extension. Section 2 describes the approach regarding the movement. Section 3 describes the approach regarding the shape. Section 4 applies the concepts in a real scenario before the final remarks.

2. Movement of regions

There are several ways to compute the distance between two locations, according to the reference surface and the adopted metrics, e.g., the length of the geodetic line or the sum of the lengths of parts of a route within a road network. Moreover, the analysis of consecutive directions of a moving point, computed as the azimuth between them, may reveal intrinsic patterns in the whole trajectory.

The concept of moving, when applied to regions, is not that clear when applied to points, except in the cases that the shape keeps unchanged, as outlined in ISO (2008): the shape may then be assumed as its centroid. Ortiz, Rocha and Medina (2020) named this scenario as the low resolution of representative trajectories. Consequently, all the patterns defined for moving points may be investigated, e.g., stops and moves, flock and encounter. The only aspect to be added is the feature orientation. The prism described in OGC (2017) is useful to represent both movement and orientation by matching the vertices of the shape along the consecutive leaves (Figure 1). However, the overall movement of a region is performed while its boundary is reshaped, i.e., the move of each infinitesimal part of the boundary does not depend on the moves of the remaining ones. Hence, the analysis of moving regions must consider changes in their shapes, e.g., scaling, rotation and increase in complexity.

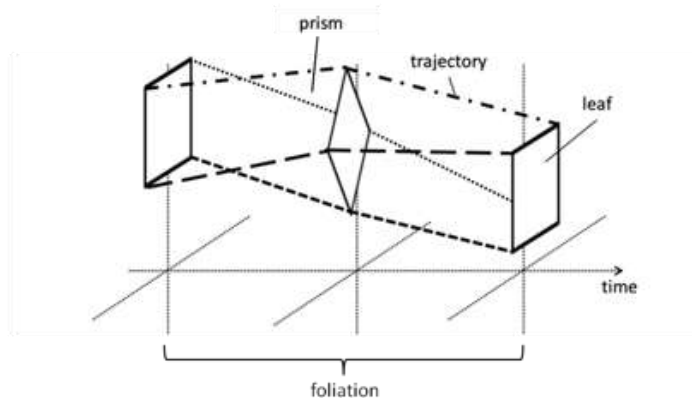


Figure 1. The changes in movement and orientation of a feature represented as a prism (OGC, 2017).

Within this context, three scenarios must be investigated: expansion or reduction of the geometry, movement, and increase or decrease of the complexity of the shape. Some approaches may be considered to analyze the geometry of moving regions.

The first approach considers each representation of a region in different instants as two sets of vertices. Hence the distance between them may be computed by either Hasseldorf, Freuchet or Jaccard distances (Deza and Deza, 2009; Alt, Knauer, and Wenk,

2004; Costa et al., 2020). Such distances do not provide any information regarding directions. It could be addressed by assuming the direction of the centroid as representative of direction.

The second approach considers the creation of trajectories based on the vertices of the consecutive leaves, as illustrated in Figure 1. This approach may be enriched by adding the trajectory composed by the centroids. In the context of this paper, such leaves may assume different shapes and, consequently, lead to ambiguations while creating the respective trajectories.

Aiming to address this issue, we propose the adoption of more simple shapes for the leaves. The first proposal is the convex hull of each leaf. However, the issue of ambiguity is only minimized because leaves continue with different quantities of vertices (see Figure 2). This issue may be addressed by assigning each vertex of the region with more vertices to the closest corner of the previous (or next) region.

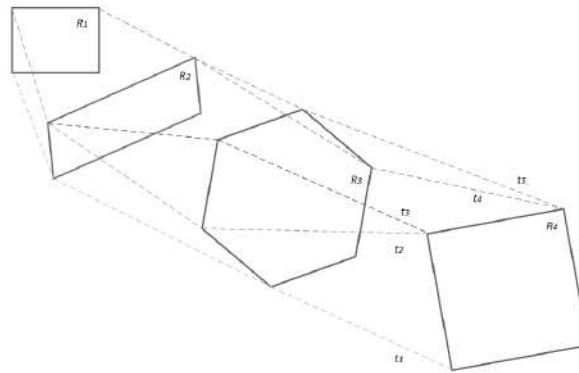


Figure 2. Example of trajectories derived from leaves of different shapes.

Junghans and Gertz (2010) describe an approach that propose the adoption of oriented minimum area bounding boxes to model each region. Such enclosing boxes are then defined by their centroid, the lengths of both the minor and the major axis, as well as the angle between the major axis and the x-axis. Later, they define a region trajectory as a set of consecutive regions R_t , as determined at time t , including the descriptions of the enclosing boxes. Analogously, the trajectories described by the corners and the centroid represent the movement of the overall geometry.

This approach may be adapted to consider the bounding box parallel to the axis, i.e., does not consider the orientation. By adopting this approach, it is possible to analyze the region movement by the joint monitoring of the a) trajectory composed by centroids; b) the trajectories composed by each vertex (towards NE, NO, SE, SO); and c) movement of the sides (expanding and movement towards N, S, E or W). Orientation could be retrieved by monitoring the directional changes observed in the vectors composed by the centroid and each vertex. Ortiz, Rocha and Medina (2020) proposed a similar method, as well as introduced the concept of resolution of the set of representative trajectories: low resolution considers only the trajectory created upon the temporal sequence of centroids; medium resolution implies the use of the vertices within each bounding box (N, S, E and O); and the high resolution consists on improving the medium resolution set by assigning vertices for the complimentary directions (NE, NO, SE, and SO).

By assigning a set of trajectories to represent the moving regions, patterns such as convoy, swarm and flock may be detected. Moreover, the asymmetric movement is supposed to be modeled based on the dissimilarities between those trajectories.

The last approach to propose is based on geometric differences. We argue that when the geometry of the region before (R_1) expands, the geometry of the region later (R_2) not covered by R_1 will return a multipolygon feature. Moreover, when the geometry of R_1 reduces, there will be parts of R_2 not covered by R_1 . On the other hand, the movement of the feature may be detected if the directions of expansion and reduction converge. The main advantage to highlight is the feasibility of detecting directional biases from the original geometries, therefore minimizing geometric uncertainties.

By observing Figure 3a, one concludes that R_1 (in black) expands east and south to assume the geometry R_2 . The direction is provided by connecting the centroids of R_1 and ($R_2 - R_1$) and by retrieving its azimuth.

By observing Figure 3b, one concludes that R_1 (in black) expands to the north and south, as well as reduces along the east and west. The expansion is given by ($R_2 - R_1$), resulting in a multipolygon composed by two polygons: north and south. Reduction is given by ($R_1 - R_2$), also resulting in a multipolygon composed by two polygons: east and west. The directions of expansion are provided by connecting the centroids of R_1 and those of each individual polygon of expansion and by retrieving each azimuth. The directions of reduction are provided by connecting the centroids of each individual polygon of expansion and of R_1 , and by retrieving each azimuth in this way it is observed that the directions diverge.

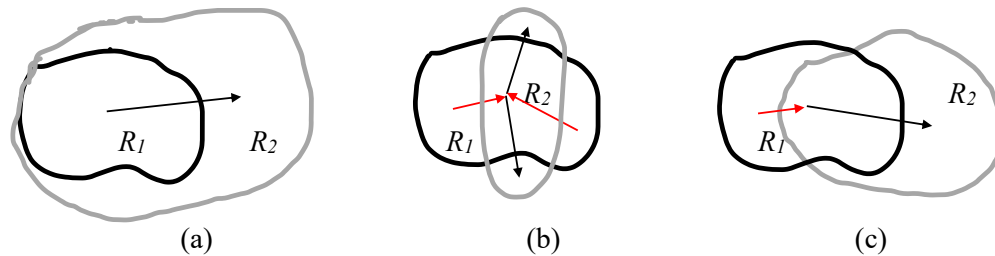


Figure 3. Different scenarios based on the difference between geometries: a) expansion; b) multipolygons for expansion and reduction; and c) feature movement.

By observing Figure 3c, one concludes that R_1 (in black) expands north and south, as well as reduces along east and west. The expansion is given by ($R_2 - R_1$), resulting in a multipolygon composed of two polygons: north and south. Reduction is given by ($R_1 - R_2$), also resulting in a multipolygon composed of two polygons: east and west. The directions of expansion are provided by connecting the centroids of R_1 and those of each individual polygon of expansion and by retrieving each azimuth. The direction of movement is provided by the same rule observed in Figure 3b. However, in this scenario, directions converge.

3. Changes in the shapes of regions

By extending the geometry of a moving object to be represented by a polygon, the trajectory will improve the intrinsic information by integrating resources of analysis of

spatial parcels. In this paper, the changes in the shapes are defined by the increase or decrease of the spatial extent covered by the geometry, and by the change in the complexity of the geometry along the timeline.

McGarigal e Marks (1995) list several well-known landscape metrics to describe different scales of landscape, according to several subject areas. For the scope of this Section, each moving region is considered an individual patch. Hence, some landscape metrics regarding the patches may adopted to analyze the dynamics of a moving region: area and shape index (or fractal dimension). Such metrics support the analysis of the temporal changes in the shape that occurred during the movement: absolute expansion or reduction of the size of an individual region and the shape complexity, compared to a standard type. Later, patterns and relationship between movement and shape could be investigated.

Duarte, Dias and Moreira (2018) proposed the use of the concept of mesh to represent the evolution of moving regions over time. Within this context, every polygon must be represented as a triangulation, and every existing triangle observed in time t_i is assigned to a correspondent triangle observed in t_{i+1} . They presented a framework to support some queries over the meshes. Such resources may be applied to complement the well-known landscape metrics to model the changes in the shapes of the moving regions.

4. Final Remarks

The objective proposed for this paper was to introduce an approach to analyze trajectories of moving regions, highlighting the aspects of displacement, direction, scale, and shape complexity.

That approach based on representative trajectories adapts the problem to be handled by well-known methods developed for moving points. The approach based on the geometric distances may enrich the analysis of moving regions by identifying directional biases from the original geometries, i.e., eliminating a source of uncertainty. Such approaches are not exclusive, i.e., they must be arranged to extract the best results, according to the application. The joint analysis of movement and shape is supposed to reveal patterns as the region moves.

Future works include the study of the interaction of moving regions, aiming to identify patterns such as encounters and avoidance and, hence detect potential merges, attractors, and avoided features. Moreover, further studies are necessary to support the process of transformation of the set of regions into the best representative set of trajectories used as the input for well-known processes of pattern mining in trajectories.

References

- Alt, H., Knauer, C., Wenk, C. (2004) “Comparison of Distance Measures for Planar Curves”, *Algorithmica*, 38. p. 45–58, doi: 10.1007/s00453-003-1042-5
- Andrienko, N., Andrienko, G., Pelekis, N., Spaccapietra. S. (2008) “Basic Concepts of Movement Data In Mobility”, In *Data Mining and Privacy: Geographic Knowledge Discovery*. Pages 15–38. Springer-Verlag Berlin Heidelberg.
- Costa, R. L. C., Miranda, E., Dias, P. and Moreira, J. (2020) “Evaluating Preprocessing and Interpolation Strategies to Create Moving Regions from Real-World Observations” in *Proceedings of the 2020 ACM Symposium on Applied Computing*, ACM 978-1-4503-6866-7 doi: 10.1145/3341105.3374019

- Deza, M. M., and Deza, E. (2009) “Encyclopedia of distances”, <http://www.uco.es/users/malfezan/Comunes/asignaturas/vision/Encyclopedia-of-distances-2009.pdf>, August.
- Duarte, J., Dias, P., Moreira, J. (2018). “A Framework for the Management of Deformable Moving Objects”. In: Mansourian, A., Pilesjö, P., Harrie, L., van Lammeren, R. (eds) *Geospatial Technologies for All. AGILE 2018. Lecture Notes in Geoinformation and Cartography*. Springer, Cham. doi: 10.1007/978-3-319-78208-9_17
- Feng, Z., Zhu, Y. (2016) "A Survey on Trajectory Data Mining: Techniques and Applications," in *IEEE Access*, vol. 4, pp. 2056-2067, doi: 10.1109/ACCESS.2016.2553681.
- ISO. (2008) ISO 19141:2008(en) - Geographic information — Schema for moving features. 2008.
- Jeung, H., Yiu, M. L., Zhou, X., Jensen, C. S., Shen, H. T. (2008) Discovery of convoys in trajectory databases. In *Proceedings of VLDB Endow. 1*, Pages 1068–1080. Doi: 10.14778/1453856.1453971
- Junghans C., Gertz., M. (2010) “Modeling and prediction of moving region trajectories”. In *Proceedings of the ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS '10)*. Pages 23–30. Association for Computing Machinery, doi: 10.1145/1878500.1878507
- Macedo, J., Vangenot, C., Othman, W., Pelekis, N., Frentzos, E., Kuijpers, B., Ntoutsis, I., Spaccapietra, S., Theodoridis, Y. (2008) “Trajectory Data Models In Mobility”, In *Data Mining and Privacy: Geographic Knowledge Discovery*. Pages 123–147. Springer-Verlag Berlin Heidelberg.
- McGarigal, K.; Marks, B. J. (1995) “FRAGSTATS: spatial pattern analysis program for quantifying landscape structure. Gen. Tech. Rep. PNW-GTR-351”. Portland, OR: U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station. 122 p.
- Mckennney, M. and Frye, R. (2015) “Generating Moving Regions from Snapshots of Complex Regions”. *ACM Trans. Spatial Algorithms Syst.* 1: 1- 30. Doi: <https://doi.org/10.1145/2774220>
- OGC. (2017) “OGC Moving Features Access”. <http://docs.opengeospatial.org/is/16-120r3/16-120r3.html>, August.
- Ortiz, A. E., Rocha, L. A., and Medina, R. J. (2020) “Modelo semántico para trayectorias de regiones em Movimento”, *Revista Ibérica de Sistemas e Tecnologias de Informação*. N° E28: 437 – 449.
- Siabato, W., Claramunt, C., Ilarri, S., and Manso-Callejo, M. Á. (2018). “A Survey of Modelling Trends in Temporal GIS”. *ACM Computing Surveys (CSUR)*, 51(2), 30.

Agrupamento de Endereços Georreferenciados para a Delimitação de Áreas Postais

Mateus R. S. Catrinque¹, Clodoveu A. Davis Jr.¹

¹Departamento de Ciência da Computação
Universidade Federal de Minas Gerais (UFMG) – Belo Horizonte, MG – Brazil

{mcatrinque, clodoveu}@dcc.ufmg.br

Abstract. *The Postal Addressing Code (CEP) is a form of sectorized representation of residential and commercial units within the Brazilian territory and is most used by logistics companies and mail services. This article aims to provide a spatial database that includes a representation of CEP areas belonging to the State of Minas Gerais based on a data clustering of data from residential and commercial addresses. From this spatial representation, the regions covered by each CEP can be materialized and used as a strategy for approximation in geocoding process.*

Resumo. *O Código de Endereçamento Postal (CEP) é uma forma de representação setorizada de unidades residenciais e comerciais dentro do território brasileiro, sendo amplamente utilizado por empresas de logística e serviços postais. Este artigo tem como objetivo fornecer uma base de dados espaciais que inclua uma representação geométrica para áreas de CEP pertencentes ao Estado de Minas Gerais a partir de um agrupamento de dados dos endereços das residências e comércios. A partir dessa representação espacial, as regiões correspondentes à área de cada CEP serão estimadas e exploradas como estratégia de aproximação no processo de geocodificação.*

1. Introdução

A tarefa de identificar um local apenas pelo endereço ou pontos de referência, por exemplo, pode ser árdua e costuma requerer cautela para evitar possíveis ambiguidades relacionadas ao nome de uma rua, bairro ou cidade dentro do território brasileiro. Segundo ([Martins et al. 2012]), é comum existir um certo grau de incerteza em relação à esses registros devido a falta de regras rígidas de padronização. Para serviços de logística e correspondências, a falta de uma boa referência de localização pode acarretar atrasos na entrega, desperdício de recursos e perda de credibilidade na prestação do serviços.

No intuito de facilitar a identificação das áreas de distribuição de correspondência a partir de um identificador único, a Empresa Brasileira de Correios e Telégrafos (ECT) desenvolveu, ao longo do território brasileiro, um sistema de codificação territorial denominado Código de Endereçamento Postal (CEP), composto por oito dígitos. A partir desses geocódigos, a ECT agrupou endereços atribuindo o mesmo código postal, de acordo com o critério de subdivisão de área que atendessem às suas necessidades logísticas.

Buscando reduzir o volume de registros necessários para se obter informações relacionadas ao CEP, esse trabalho propõe uma base de dados contendo representações

geométricas para as áreas de endereçamento postal relacionadas ao Estado de Minas Gerais. Para a delimitação das geometrias postais, os dados de residências e comércios provenientes das bases do OpenAddresses e BHMAP foram agrupados de acordo com o seu atributo postal. Os *clusters* espaciais gerados, foram submetidos a duas abordagens de envoltório distintas, o *Convex-Hull* e o *Concave-Hull*, sendo posteriormente avaliado qual das estratégias de geração de limites para as geometrias de CEP apresentou menor ocorrência de sobreposição de áreas.

A criação de uma base aberta contendo dados de áreas postais capaz de prover resultados para buscas por localização de CEP com o retorno de um único registro relacionado ao alvo, além de seu ganho operacional para essa finalidade, pode também ser apontada como mais uma possibilidade de recurso alternativo frente à base proprietária da ECT, que apesar de ser empresa pública, optou por comercializar o geocódigo em detrimento à disponibilização gratuita dos dados para fomentar a pesquisa e desenvolvimento do setor de logística nacional.

Os fundamentos teóricos que conceituam essa proposta, são explicitados ao longo da Seção 2. Na Seção 3, o processo para a construção do *cluster*, bem como os refinamentos para obter uma geometria aproximada são detalhados. Por fim, nas seções 4 e 5, são dispostos argumentos sobre os resultados e apresentadas as conclusões e trabalhos futuros pretendidos a partir do projeto desenvolvido.

2. Trabalhos Relacionados

O conceito de geocodificação de endereços é descrito em [Davis and Alencar 2011] como aquele em a partir de um endereço postal se obtém uma localização geográfica determinada. Esse mesmo conceito pode ser visto de forma mais ampla como a localização de qualquer indicador que possibilite referenciar um lugar de interesse. Essa definição pode ser observada em [Martins et al. 2012], onde também é reconhecida a importância dos códigos postais como um dos principais recursos para transmitir uma posição geográfica.

Quando se observa unidades residenciais e comerciais, é possível notar registros que comungam dos mesmos atributos. Ao processo de criação de um agrupamento de entidades que possuem alguma semelhança entre si é dado o nome de *clustering*. Segundo [Kolatch et al. 2001], em um conjunto de dados, o *clustering* permite a criação de grupos de objetos organizados de acordo com suas similaridades que, em dados geocodificados, possibilita uma generalização de componentes espaciais.

Para delinear o contorno das representações espaciais oriundas dos *clusters* obtidos, abordagens como o *convex hull* e *concave hull* são meios de estimar o conjunto de pontos de um agrupamento. No entanto, vale frisar que, conforme observado em [Moreira and Santos 2007], tais modelagens, por mais aproximadas que sejam, ainda podem resultar em distorções, em especial quando utilizadas abordagens como a envoltória côncava, que não possui apenas uma forma resultante, pois depende tanto da abordagem, quanto do número de k vizinhos definido no momento de sua implementação.

Ao realizar o procedimento similar para gerar representações espaciais dos endereços postais de uma cidade holandesa, [Penninga et al. 2003] aponta que quando os dados de entrada fornecidos pelas bases exploradas são discriminados na forma de pontos, a geometria das áreas obtida pelo agrupamento de pontos por código postal pode resul-

tar em distorções, devido à essa forma de representação ser apenas uma aproximação do centróide dos componentes do *cluster* espacial gerado.

3. Metodologia

Diante do objetivo do projeto de compreender a distribuição das unidades residenciais e comerciais dentro do território de Minas Gerais, especificamente em Belo Horizonte, foram exploradas as bases georreferenciadas disponíveis em dois repositórios: uma no OpenAddresses¹ e a segunda, a base de dados aberta oficial fornecida pela Prefeitura de Belo Horizonte (PBH)².

Para facilitar no entendimento do desenvolvimento da base de dados de áreas de CEP, segue na Figura 1, o fluxo de trabalho com as etapas de execução do processo:

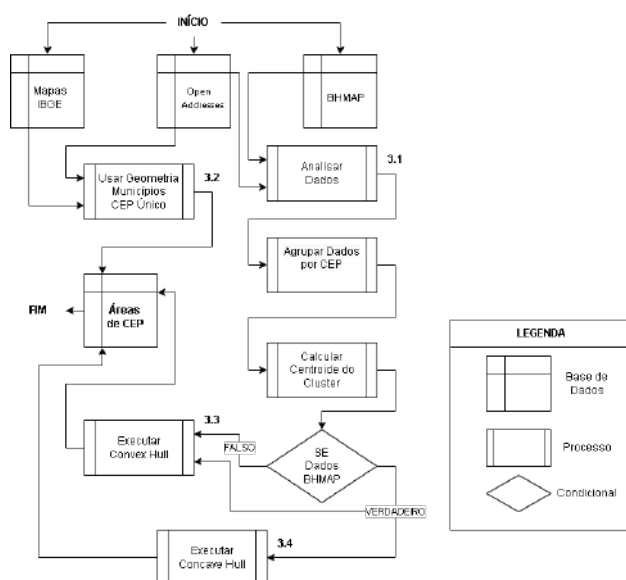


Figura 1. Diagrama de Atividades

3.1. Análise de dados

A partir dos registros explorados, foi possível observar o comportamento e as regiões com maior concentração de dados, evidenciando o grande volume de dados relacionados à região metropolitana da capital mineira dispostos nas bases, sendo a regiões em vermelho as com maior concentração.

3.2. CEP Único Municipal

Com o conhecimento sobre o volume e as áreas de maior concentração de dados, buscou-se verificar a relação entre a quantidade de CEPs que cada município de Minas Gerais abrangia.

Observou-se que mais de 500 municípios presentes na base do OpenAddresses, tem seus endereços associados a um único CEP. Para esses casos, o formato ideal da

¹<http://openaddresses.io/>

²<http://bhmap.pbh.gov.br/v2/home.html>

área de CEP equivale ao próprio perímetro municipal. Portanto, para esses casos, foram utilizados os limites municipais disponibilizados pelo Portal de Mapas do IBGE³, associando-os ao CEP correspondente.

Para os demais casos, foi realizada uma operação de agrupamento de endereços, subdividindo os dados em conjuntos formados a partir do atributo CEP.

3.3. Definição das geometrias

A partir do agrupamento, foi executada uma técnica de *convex hull* para os casos em que o município não possuía CEP único, gerando uma estimativa geométrica baseada na criação de arestas interligando os pontos mais extremos de cada conjunto formado de acordo com sua proximidade, obedecendo o limite de duas ligações por vértice.

Como forma de selecionar os vértices da geometria de área de cada CEP, foi estimado um centro geométrico para cada agrupamento e calculada a distância média entre os endereços georreferenciados do grupo para o centroide. A partir dessa operação, desconsiderou-se os *outliers* no processo de escolha dos vértices para delinear o polígono representante da área de CEP.

3.4. Aprimoramento da delimitação de área

O comportamento para as geometrias de CEP de regiões com alta concentração de unidades se mostrou extremamente irregular. Em Belo Horizonte, observa-se na figura 2a a formação de geometrias com áreas extensas cobrindo grande parte do território da capital e também o excesso de sobreposições entre os polígonos.

Na tentativa de suavizar essas distorções de representação, partiu-se para uma remodelagem dos polígonos, privilegiando a manutenção dos possíveis contornos que um agrupamento de pontos geoespaciais comumente possui, conforme pode ser visto na figura 2b

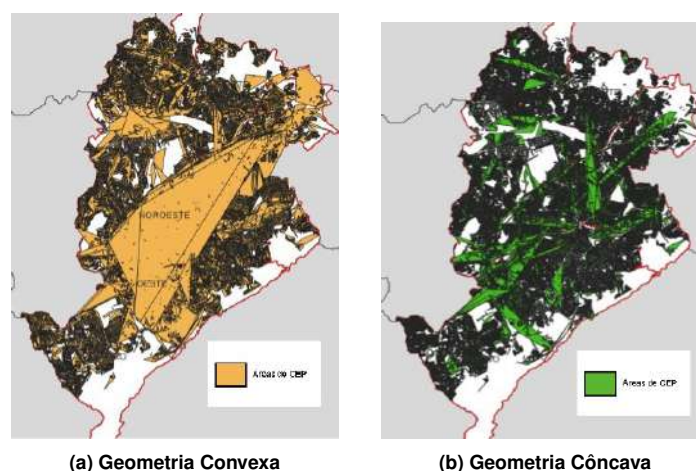


Figura 2. Comparação entre geometrias de CEP para Belo Horizonte

Um modelo que se mostrou eficiente em atender esses requisitos foi o *Concave Hull* baseado nos *K-Nearest Neighbors* (KNN), visto em [Moreira and Santos 2007],

³<https://portaldemapas.ibge.gov.br>

onde o número de vizinhos foi definido arbitrariamente em 10 com devido volume de registros mínimo encontrado na formação do menor *cluster* de CEP contido na base de dados. Ele tem como diferencial em relação ao *Convex Hull* o fato de explorar apenas os vizinhos próximos ao ponto de cálculo em que se encontra, isso é, ao chegar em um determinado ponto, apenas suas vizinhanças mais próximas serão calculadas, para então partir para o próximo e assim, sucessivamente.

Após a aplicação de uma delimitação de envoltório côncavo como alternativa, os contornos de área ganharam melhor definição e a forma dos polígonos foi reduzida, quando comparado aos gerados pela abordagem anterior. Essa diferença também pode ser notada na Figura 3, ao se observar a redução de distorções na área das geometrias geradas para os CEPs citados abaixo:

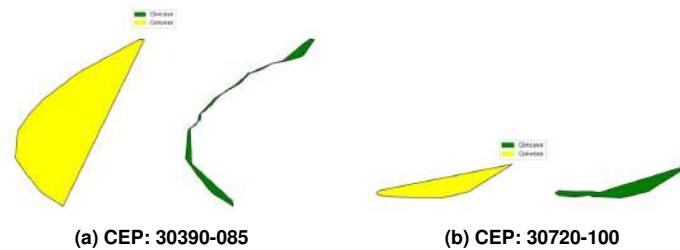


Figura 3. Comparação entre as geometrias geradas para os CEPs

4. Resultados e Discussão

Ao final do processo de delimitação de geometrias para as áreas de CEP, uma base de dados espaciais foi gerada a partir do resultado de três operações chave, sendo elas: o uso dos limites municipais das cidades com CEP único como polígono para representar suas áreas de código postal; a operação que estimou o polígono de cada um dos CEPs restantes baseando-se em uma envoltória convexa; e a remodelagem côncava proposta para os agrupamentos postais, a partir dos registros advindos da base do BHMap em conjunto com os do OpenAddresses.

Dentre os aspectos positivos dessa nova base gerada destaca-se a proposta de uma representação para as áreas de CEP, a redução no volume de registros necessários para obter informações relacionadas ao endereçamento postal de 7.418.399 (somando os registros do OpenAddresses e BHMap) para 38.171 e a simplificação no resultado das consultas graças a possibilidade de retorno de um único registro representante nas operações relacionadas ao CEP realizada na base.

Os atributos que compõem a base proposta, estão descritos na figura 4.

Coluna	Descrição	Tipo
CEP	Código de Endereçamento Postal	primary key(string)
Cidade	Cidade a qual o código pertence	string
QTD Unidades	Quantidade de endereços para gerar a área de CEP	integer
geometry	Geometria espacial do CEP	geometry

Figura 4. Descrição do atributos na base de dados de CEP gerada

Um dos principais motivadores da proposta é o de prover dados relacionados aos CEPs, contrapondo a política da ECT que se recusa a liberar o acesso aberto ao seu catálogo. No intuito de fomentar pesquisas relacionadas aos códigos postais e possibilitar o acesso aos dados de CEP do Estado de Minas Gerais, a base de dados gerada e o resultado das análises exploratórias estão disponíveis para uso e consulta no seguinte repositório:

- <https://github.com/mcatrinque/areas-cep-mg>

5. Conclusões

A partir do agrupamento de endereços contidos nas bases de dados exploradas, foram geradas representações espaciais para as áreas de CEP do Estado de Minas Gerais. Observando a relação entre a forma dos CEPs em cidades e a alta concentração de *clusters* postais, constatou-se a suavização dos contornos das geometrias postais pode facilitar na visualização e precisão da estimativa de área do CEP, além de diminuir a ocorrência de sobreposições entre as áreas.

A abordagem proposta focou em explorar a criação de geometrias a partir das demarcações de endereço por meio de pontos de referência no espaço, sendo essa a forma amplamente disponibilizada nas bases de dados brasileiras. Esse fator é a principal diferença em relação a abordagem disposta em [Penninga et al. 2003], que utilizou os limites geométricos das propriedades da cidade de Zevenhuizen, disponibilizados pelo sistema de cadastro público holandês. Para garantir uma maior aplicabilidade desse tipo de exploração nas bases de dados fornecidas pelo governo brasileiro, foi necessário abrir mão da precisão no processo de formação de geometrias de CEP.

A respeito da utilização do CEP como referência espacial, observou-se que o identificador já não reflete a mesma padronização da época de sua implantação. O crescimento populacional, a diferença entre as geometrias de CEP formadas e a imprecisão causada pelas interseções entre áreas, aponta para a necessidade de novas alternativas de divisão espacial. Como trabalhos futuros, pretende-se apresentar um modelo de geocódigo que atenda a necessidade de uma nova subdivisão nacional em áreas de entrega.

Referências

- Davis, C. A. and Alencar, R. O. (2011). Evaluation of the quality of an online geocoding resource in the context of a large brazilian city. *Transactions in GIS*, 15(6):851–868.
- Kolatch, E. et al. (2001). Clustering algorithms for spatial databases: A survey. *PDF is available on the Web*, pages 1–22.
- Martins, D., Davis Jr, C. A., and Fonseca, F. T. (2012). Geocodificação de endereços urbanos com indicação de qualidade. *Proceedings XIII GEOINFO*, pages 36–41.
- Moreira, A. and Santos, M. (2007). Concave hull: A k-nearest neighbours approach for the computation of the region occupied by a set of points. *GRAPP*, pages 61–68.
- Penninga, F., Verbree, E., Quak, W., and van Oosterom, P. (2003). Construction of the planar partition postal code map based on cadastral registration. *ACM International Symposium on Advances in GIS*, page 134–140.

Contribuição de produtos de área queimada na Amazônia Maranhense: proposta de avaliações combinadas

Paulo Henrique A. Leão¹, Liana O. Anderson¹, Ana Larissa R. de Freitas¹, João Bosco Coura dos Reis¹, Swanni T. Alvarado^{2,3}, Celso Henrique L. Silva-Junior⁴, Tiago Massi Ferraz³

¹Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden), São José dos Campos - SP - Brazil

²Facultad de Ciencias Naturales. Universidad del Rosario - Bogotá - Colombia

³ Programa de Pós-graduação em Agricultura e Ambiente, Universidade Estadual do Maranhão, Praça Gonçalves Dias, CEP 65800-000, Balsas, Maranhão, Brazil

⁴University of California, Los Angeles - CA - Estados Unidos

henri.leaos@gmail.com, liana.anderson@gmail.com,
alarisig@gmail.com, joaodosreis89@gmail.com, swanni_ta@yahoo.es,
celsohlsj@gmail.com, ferraztm@gmail.com

Abstract. *The fusion of fire products can significantly refine the identification of burned areas, and based on that, we evaluated the possibility of using three products, MCD64A1, GABAM and Fire_CCI, for the years 2012 to 2019 in the Maranhão Amazon region. We succeeded in increasing the identification of the burned area of 24%, reaching up to 50% more than the MCD64A1, a product that would have the best coverage result for the region.*

Resumo. *A fusão de produtos de fogo podem refinar significativamente a identificação de áreas queimadas, e partindo disso avaliamos a possibilidade de uso de três produtos, MCD64A1, GABAM e Fire_CCI, para os anos de 2012 a 2019 na Amazônia maranhense. Conseguimos trazer aumentos na identificação da área queimada de 24% podendo chegar até 50% a mais que o MCD64A1, produto que teria o melhor resultado de cobertura para a região.*

1. Introdução

A floresta amazônica é a maior floresta tropical do mundo, cobrindo quase 7 milhões km² e se estendendo por nove países da América do Sul. Destes nove, o Brasil detém a maior área contínua de floresta na região, cerca de 60% da bacia amazônica distribuída em nove estados federativos [RAISG, 2015].

O Maranhão originalmente tinha uma cobertura florestal de 110.400 km² (cerca de 33% do estado) [Silva-Junior *et al.*, 2022]. Porém, desde meados da década de 60, o desmatamento em grande escala voltado principalmente para destinar áreas para agricultura e pecuária, além da ocorrência de queimadas e incêndios florestais intensificados devido à fragmentação florestal [Silva-Junior *et al.*, 2018], resultou na perda significativa da cobertura original da floresta amazônica maranhense: para 25% (24.700 km²) em 2016 e de 24 % (23.967 km²) em 2019 [Silva-Junior *et al.*, 2020].

Além disso, visto que o fogo gerado de forma natural é raro no bioma amazônico, quase todos os incêndios florestais atuais são provocados por queimadas deliberadas, acidentais ou intencionais, que saem do controle e se tornam um incêndio florestal [Bush 2008]. Esses incêndios florestais contribuem significativamente para a mudança climática global, consumindo biomassa vegetal e transferindo parte do estoque de carbono (C) associado para a atmosfera [Pessoa *et al.*, 2020; Silva *et al.*, 2020].

Pontuados na preocupação de identificar e quantificar áreas queimadas, diversas abordagens metodológicas têm sido desenvolvidas utilizando aplicações de sensoriamento remoto para detecção e monitoramento de queimadas e incêndios florestais. A diversidade de metodologias, aliada à disponibilidade de múltiplos sensores, e o rápido desenvolvimento de novas tecnologias reflete o elevado número de produtos de área queimada [Pessôa *et al.*, 2020].

Pessôa *et al.* (2020) demonstra que produtos globais com uma resolução mais grossa e conservadora usados em conjunto com produtos em escala regional ou produtos com uma menor resolução espacial, podem refinar significativamente a identificação de áreas queimadas e, conseqüentemente, o cálculo das emissões de carbono provenientes do fogo. O processo de escolha pode envolver a fusão de mais de um produto para otimizar suas vantagens e produzir dados mais consistentes para as necessidades do usuário, aproximando-se da verdadeira área total queimada e sua distribuição regional [Pessôa *et al.* 2020].

Partindo dessa premissa, foi avaliado neste trabalho a possibilidade de uso de três produtos de área queimada, MCD64A1, GABAM e Fire_CCI, para os anos de 2012 a 2019 na amazônia maranhense. O intuito foi de entender as diferenças e ganhos de informações para a formulação de uma única informação de área queimada mais consistente.

2. Área de estudo

O estado do Maranhão está localizado na região nordeste do Brasil, e dentro do estado possui 3 biomas: Amazônia, Cerrado e Caatinga. A área de estudo do trabalho corresponde à floresta amazônica do estado do Maranhão. Da cobertura florestal remanescente (23.967 km²) na região em 2019, mais de 70% estão contidas dentro de Unidades de Conservação (UC) e Terras Indígenas (TI), como as TI Caru, Alto Turiaçu, Araribóia e Awá ou a Reserva Biológica do Gurupi [Silva-Junior *et al.*, 2022].

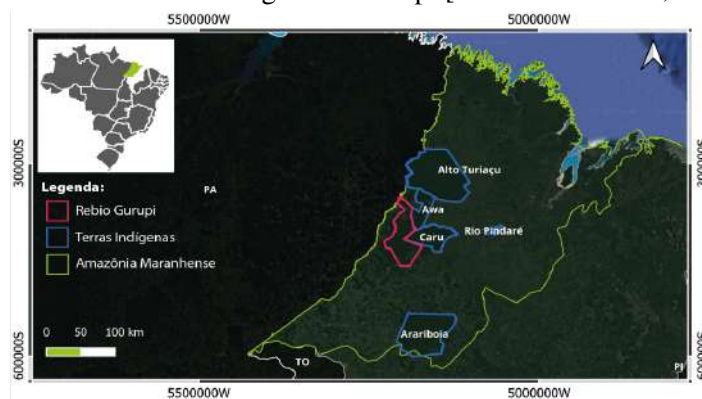


Figura 1. Amazônia Maranhense.

3. Metodologia

Três produtos globais amplamente citados na literatura foram utilizados para o recorte da Amazônia Maranhense nesta pesquisa (Tabela 1), para o período de 2012 a 2019, conforme a disponibilidade de dados. A partir dessas informações, foi analisado a distribuição espacial e as sobreposições dos dados, identificados como: MCD64A1; MCD64A1 + Fire_CCI; Fire_CCI; Fire_CCI + GABAM; GABAM; GABAM + MCD64A1 e Todos os produtos [Pessôa *et al.*, 2020].

Tabela 1. Produtos de área queimada selecionados.

Produto	Satélite	Resolução espacial	Fonte
MCD64A1	MODIS	500m	NASA
Fire_CCI	MODIS	250m	ESA
GABAM	Landsat 8	30m	Long T. (2019)

Para verificar a abrangência da área queimada, foram utilizados os dados de focos de calor do Programa Queimadas - INPE [INPE, 2022], para quantificá-los dentro das classes apresentadas e em seu entorno. A avaliação permite confirmar a existência de fogo em um elemento de resolução da imagem (pixel). Assim, a adição de novos produtos para compor uma área queimada pode abranger uma maior distribuição frente à ocorrência de fogo, do que um único produto poderia identificar [Freitas, 2021].

4. Discussão e Resultados

Inicialmente foram avaliadas as diferenças de área queimada entre os produtos, para um entendimento melhor sobre as possíveis contribuições de cada um. Ao mesmo tempo em que foi possível entender o comportamento das ocorrências de áreas queimadas ao longo do tempo. Os anos de 2012, 2015 e 2017 foram os anos de maior área queimada para os três produtos. Contudo, os produtos diferem no total de área queimada mapeada, como explícito na Figura 2. O produto MCD64A1 mapeou os maiores valores de área queimada em todos os anos da série temporal (entre 446km² em 2018 e 9.738 km² em 2015). Seguido pelo GABAM que identificou entre 281km² em 2018 e 4.237 km² em 2015. O produto que identificou as menores áreas foi o Fire_CCI (entre 260km² em 2018 e 2.202 km² em 2012).

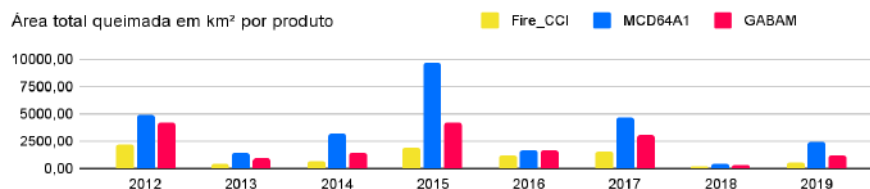


Figura 2. Tamanho de áreas queimada em Km² dos produtos Fire_CCI, Gabam e MCD64A1 para a amazônia maranhense no período de 2012 a 2019.

As diferenças entre os produtos ainda são constatadas quando relacionados aos dados de focos de calor contidos dentro da área de queima de cada produto. Na Figura 3 é possível perceber que mesmo o produto GABAM não sendo aquele que identificou as maiores taxas de área queimada, ainda sim é o produto que mais conseguiu identificar áreas com focos de calor. Como citado por Pessoa (2020), o produto GABAM apresenta muitas manchas de áreas queimadas, o que reflete o maior número de células com baixa proporção de queima. Então mesmo com menos áreas queimadas, o produto consegue alcançar com um trato mais fino áreas menores não identificadas por outros produtos com maior resolução espacial.

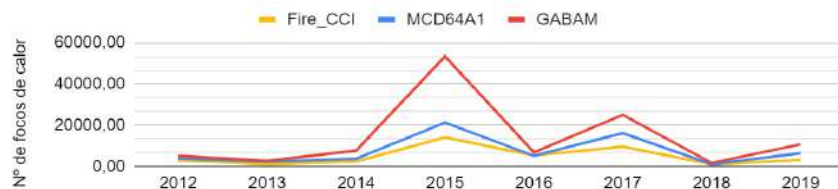


Figura 3. Relação entre o número de focos de calor/produto de área queimada.

Essa informação fica evidente na Figura 4, os pontos de focos de calor identificam possíveis áreas queimadas que não foram identificadas em cada produto separado. Mas essas informações podem ser somadas para complementar umas às outras, de forma a obter um mapeamento de área queimada com maior representação das áreas queimadas e mais condizente com a realidade.

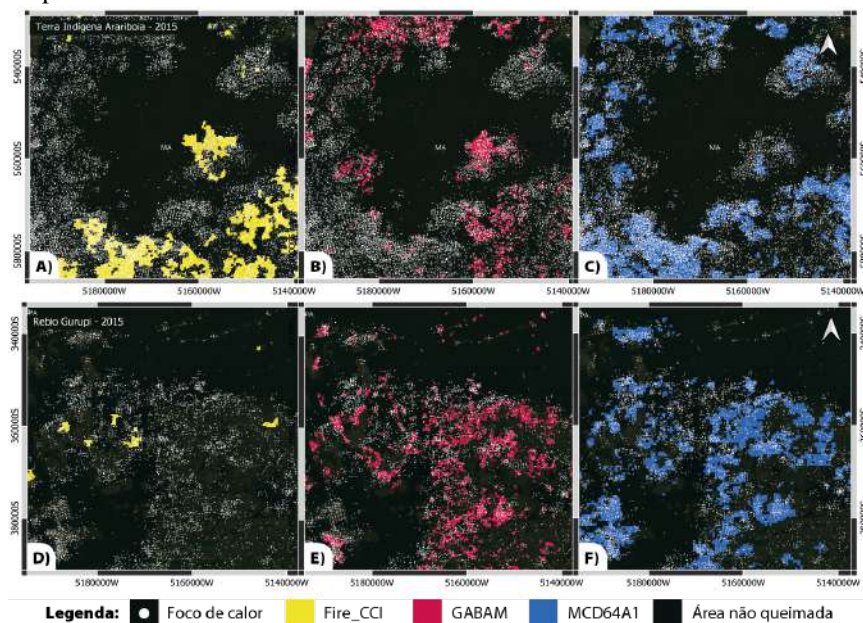


Figura 4. Diferenças nos resultados dos produtos nas áreas de terra indígena Araribóia (A, B e C) e na Reserva Biológica do Gurupi (D, E e F) em comparação com o pontos de foco de calor do Programa Queimadas.

O mesmo comportamento é encontrado em todos os anos analisados (Figura 5). O produto que mais contribuiu com pontuação única no produto final foi o MCD64A1, com valores que variam de 30% a 56% durante o período selecionado. Seguido do GABAM com contribuições únicas que variam entre 17% a 32% e por último o Fire_CCI, com as menores contribuições para área final que variam entre 7% a 22% (este último sendo um evento muito fora da média de porcentagens anuais que é de 11%). Em média, comparando os tamanhos de área queimada dos produtos com o produto final, ele adicionou 75% mais informação de área queimada que o Fire_CCI (entre 79 a 84% durante os anos), 39% mais informação que o MCD64A1 (entre 24 a 50% entre os anos) e 62% mais informação que o GABAM (entre 50 a 69% entre os anos).

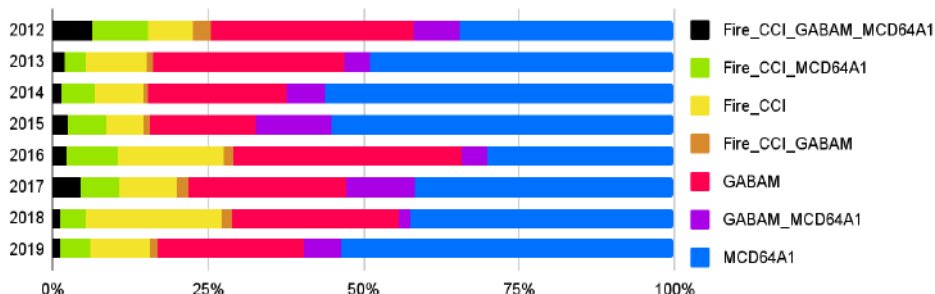


Figura 5. Classes de contribuição para o produto final de área queimada.

Essas contribuições ficam mais evidentes quando visualizadas espacialmente, para entender como as classes se comportam ao mesclar produtos diferentes para a mesma área (Figura 6). É possível perceber as contribuições únicas de cada produto, como a concentração de áreas do Fire_CCI mais ao sul (Figura 6C), próximo a TI Araribóia. Também fica perceptível a concentração do MCD64A1 mais ao norte (Figura 6A) e GABAM no centro do estado (Figura 6B). Esses agrupamentos são uma possível resposta ao gradiente de diferentes vegetações do estado e as mudanças das ecorregiões ali presentes. Se faz necessário mais estudos para um melhor entendimento.

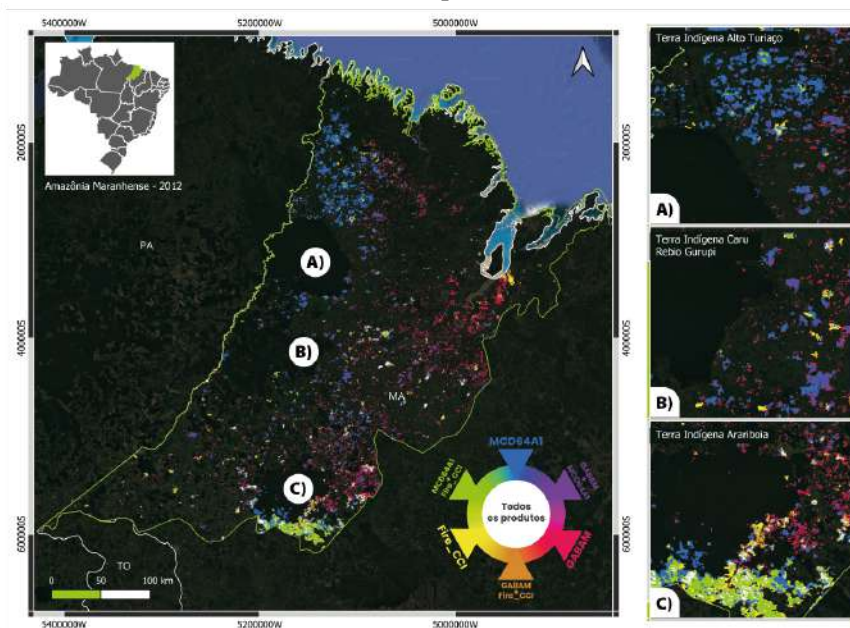


Figura 6. Representação espacial e classificada do produto final de área queimada para o ano de 2012. Enfoque nas Ti Alto Turição, Caru e Araribóia (A, B e C) e na Rebio Gurupi (B).

A contagem de focos na Figura 7 também evidencia e valida a existência de uma queimada nas classes definidas e confirmam os comportamentos para 2015 e 2017. A exceção foi em 2012, já que 2014 e 2019 apresentaram mais focos de calor, porém com menos áreas queimadas identificadas em todos os produtos.

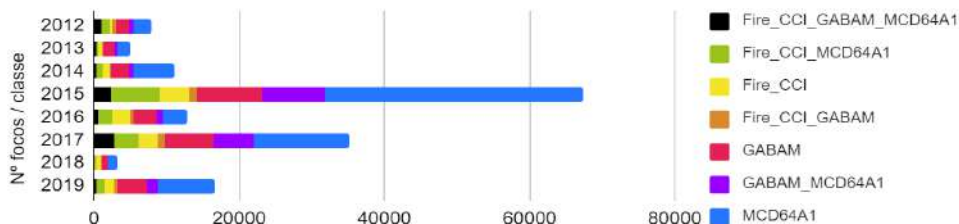


Figura 7. Número de focos dentro das classes.

Dentro do produto final desenvolvido, a classe que apresentou maior número de focos de calor sobrepostos foi a de MCD64A1 com uma contagem máxima de 35.606 ocorrências, seguido pela classe GABAM com 8873 focos e a classe de interseção GABAM_MCD64A1 com 8804 focos, todos para o ano de 2017. Um ponto interessante foi perceber a classe que compõe a concatenação entre todos os produtos, que é uma das

menores em cada ano, tanto em quantidade de focos de calor (Figura 7), quanto em tamanho de área de contribuição (Figura 5), enfatizando que os produtos abrangem locais diferentes entre si, e poucas foram mapeadas por todos os produtos.

7. Conclusão

A partir das análises realizadas neste trabalho, pode-se inferir que a combinação de diferentes produtos de área queimada, podem de fato trazer uma melhor identificação destas áreas, no que diz respeito a amazônia maranhense. Houve um aumento na identificação da área queimada de 24%, podendo chegar até 50% a mais que no produto que mais identificou áreas queimadas para a região, o MCD64A1. Esses dados correlacionados com os focos de calor de cada ano, demonstram que a área queimada identificada no produto final condiz com os focos de calor observados pelos diferentes sensores no período selecionado.

Agradecimentos

Ao Projeto MAP-Fire Maranhão (FAPEMA/FAPESP processos nº 2020/16457-3 e Nº CACD 02989/20) e pela bolsa FAPEMA processo nº 2022/06158-4.

Referencias

- Bush, M., *et al.* (2008) . “Fogo, mudanças climáticas e biodiversidade na Amazônia: uma perspectiva do Holoceno Tardio”. *Philos. Trans. R. Soc. B Biol. Sci.*, 363.
- Freitas, A. L. R. Exposição a incêndios florestais em unidades de conservação de uso sustentável na Amazônia. 2021. 134 p. Dissertação (Mestrado em Sensoriamento Remoto) - INPE, São José dos Campos, 2021.
- INPE, (2022). “Programa Queimadas - Perguntas frequentes”. Disponível em: <http://www.inpe.br/queimadas/portal/informacoes/perguntas-frequentes>.
- Pessôa, A.C.M., *et al.* (2020) “Intercomparison of Burned Area Products and Its Implication for Carbon Emission Estimations in the Amazon”. *Remote Sens.* 2020.
- RAISG - Rede Amazônica de Informações Socioambientais Georreferenciadas Desmatamento na Amazônia (1970–2013). Disponível online: <https://www.amazoniasocioambiental.org/> (29/08/2022).
- Silva-Junior C.H.L., *et al.* (2018). “Drought-induced Amazonian wildfires instigate a decadal-scale disruption of forest carbon dynamics”. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2018, 373 20180043.
- Silva-Junior C.H.L., *et al.* (2020). “Amazon forest on the edge of collapse in the Maranhão State, Brazil”. *Land Use Policy*, Volume 97.
- Silva-Junior C.H.L., *et al.* (2022). “Forest Fragmentation and Fires in the Eastern Brazilian Amazon-Maranhão State, Brazil”. *Fire*, 5, 77.
- Silva C.V.J., *et al.* (2020). “Estimating the multi-decadal carbon deficit of burned Amazonian forests”. *Environmental Research Letters*, 15 114023.
- Long, T., *et al.* (2019). “30 m Resolution Global Annual Burned Area Mapping Based on Landsat Images and Google Earth Engine”. *Remote Sens.*, 11, 489.

The effects of COVID-19 on Chinese commuting patterns in early 2020

Vander L. S. Freitas¹, Leonardo B. L. Santos^{2,3}, Ming Tang⁴,
Yong Zou⁴, Elbert E. N. Macau⁵

¹Department of Computing – Federal University of Ouro Preto (UFOP)
Ouro Preto – MG – Brazil

²National Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)
Sao Jose dos Campos – SP – Brazil

³National Institute of Space Research (INPE)
Sao Jose dos Campos – SP – Brazil

⁴School of Physics and Electronic Science – East China Normal University
Shanghai – China

⁵Institute of Science and Technology – Federal University of Sao Paulo (UNIFESP)
Sao Jose dos Campos – SP – Brazil.

vander.freitas@ufop.edu.br, santoslbl@gmail.com, tangminghan007@gmail.com,
yzou@phy.ecnu.edu.cn, elbert.macau@unifesp.br

Abstract. *We build mobility networks from Chinese commuting data and track network metrics for the two months before the WHO pandemic announcement. The Wuhan travel ban on 23 January imposed changes to the level of importance of some central cities in the commuting patterns. While Beijing was the most important city in both the inflows and outflows, Wuhan and other cities became more relevant after the transition.*

1. Introduction

The World Health Organization (WHO) declared the COVID-19 pandemic on 11 March 2020, after the cases of the new Coronavirus increased about 13-fold outside China [Cucinotta and Vanelli 2020, Li et al. 2020]. Wuhan (province of Hubei) is the city of patient zero, whose notification date back to 11 January 2020, followed by cases in Beijing, Shenzhen, and Shanghai in the following days. As of 10 September 2022, China had confirmed about 2,615,600 cases and 15,000 deaths [Dong et al. 2020].

Evidences suggest that the people commuting patterns are behind the spread of viral diseases [Freitas et al. 2020a, Kraemer et al. 2020], and the outbreaks might trigger individual and governmental-induced mobility restrictions [Freitas et al. 2020b]. It is of paramount importance to quantify and qualify the changes in mobility patterns during such events.

The Complex Network approach emerges as a natural mechanism to handle mobility data, taking areas as nodes and movements between origins and destinations as edges [Santos et al. 2019]. A Complex Network is a graph, a set of nodes (elements) and edges (relations between elements) representing a Complex System: a system with several elements with non-trivial relations.

In this paper¹, we build several networks with the commuting of each day between 01 January 2020 and 29 February 2020. We monitor how their topological indexes change over time, emphasizing four important dates: the beginning of the Spring festival travel rush (*Chunyun* - 10 January), the Wuhan travel ban (23 January), the Lunar new year (25 January), and the end of the Spring Festival travel rush (18 February) [Tan et al. 2021].

Our results show that the travel ban imposed changes in Chinese commuting patterns. While before the restrictions Beijing was the most important in both the in and outflows of people between cities, Wuhan and other central cities assumed an important role in the following days. We also observe this transition from shortest paths-based network metrics.

This paper is organized as follows: Section 2 presents the methodology, followed by the results and discussion of Section 3. The conclusions are in Section 4.

2. Methodology

2.1. Mobility and geographical data

We use the Baidu Mobility Data [China Data Lab 2020, Hu et al. 2020] from Jan 1 to Feb 29, 2020, two months prior to the WHO pandemic announcement. The dataset contains the daily inflows and outflows of people between origins and destinations within 340 Chinese cities. The inflows and outflows compose two separate adjacency matrices. The former presents the percentage of people that moves from i to j , concerning j . Node j tracks the relative inflows from every other node and its inflows sum up to one. The outflows behave similarly, but with the distribution of outflows concerning j . Moreover, each node has at most 100 neighbors.

As a preprocessing step, we remove from the analysis the cities with unavailable city codes and geographical coordinates [China Data Lab 2020, Lab 2020] and end up with the daily flows between 303 cities over the aforementioned 60 days.

2.2. Network analysis

Based on mobility data, we build 60 networks, where nodes are places and connections are flows of people commuting from one place to another within a day.

A network is a graph $G(V, E)$ with $N = |V|$ nodes and $L = |E|$ links. We use here weighted networks, whose link weights W_{ij} between pairs of nodes i and j are the flows [Barabási and Pósfai 2016].

We characterize the networks through their average weighted betweenness b_w , weighted closeness c_w , and diameter l_{max} .

Since the flows of people are the weights in the network representation, the concept of distance when computing the shortest paths must have a different interpretation: the higher the weights between two nodes, the closer they are. Thus, we use the inverse of link weights to compute the shortest path lengths l_{ij} .

The weighted betweenness quantifies the node's importance in the shortest paths of the network. The higher its value, the more shortest paths pass through it. The weighted

¹Data and source code are available at: https://github.com/vanderfreitas/chinese_commuting_patterns

closeness states how close a node is to the others. Mathematically, it is the inverse of the average shortest path lengths between i and the other nodes. The network diameter is the highest shortest path, or simply the highest possible distance (geodesic) between any pairs of nodes.

3. Results and discussion

Figure 1 shows the mobility networks with the inflows and outflows of 2 January 2020 and 2 February 2020. The node colors correspond to the weighted betweenness and edge colors represent the weights (flows). Note that the weights are in terms of percentages and have comparable magnitudes in the four subfigures. What changes from one subfigure to another is their disposition. Thus, one cannot infer the number of travelers, but the rates of incoming/outgoing people from/to point A to/from point B.

Regarding the inflow networks, the higher betweenness centralities emerge in cities that contribute more by sending people to other places, since they monopolize the percentages and serve as a distribution hub. On the other hand, those with higher b_w in the outflow networks are preferred destinations, because they receive a higher percentage of people that leaves other cities.

Figure 1a and 1b shows that Beijing fits the role of a node that is important in the shortest paths for distributing and receiving flows of people, before the travel ban. The situation changes after the travel ban, with Wuhan becoming an important local bridge within the Hubei province. About 11 cities from the same province received more than 20% of their incoming flow from Wuhan in the period. It does not mean that more people commuted from Wuhan, but that most travelers that reached the mentioned cities came from Wuhan.

Figure 2 presents the time series of a few selected network metrics: weighted betweenness b_w , weighted closeness c_w , and diameter l_{max} . The latter is inherently global whilst the others are node-wise, which justifies showing their average. We decided to use only weighted network metrics and, although others present transitions in the period of the travel ban, their amplitudes are comparable and not worth showing.

All metrics present a transition between the Wuhan travel ban on 23 January, and the Lunar new year, on the 25th, especially in the inflow networks. Recall the differences between nodes of Figure 1a and 1c, that correspond to the inflow networks of before and after the travel ban. The nodes of the first seem to have similar importance in the shortest paths (Beijing is an exception with higher b_w than the others). In contrast, in the second network one sees more nodes with higher importance (redder colors), increasing the average. Interestingly, Beijing is not in the top three anymore, in agreement with [Tan et al. 2021], that states that the outflow from the big cities decreased faster than others in the transition. The weighted closeness has a similar behavior as observed with b_w and the diameter follows the trends observed in b_w and c_w , although in a more wiggling fashion.

4. Conclusions

We discuss the changes in the Chinese commuting network during the first two months of 2020, before the COVID-19 announcement by the World Health Organization. The data

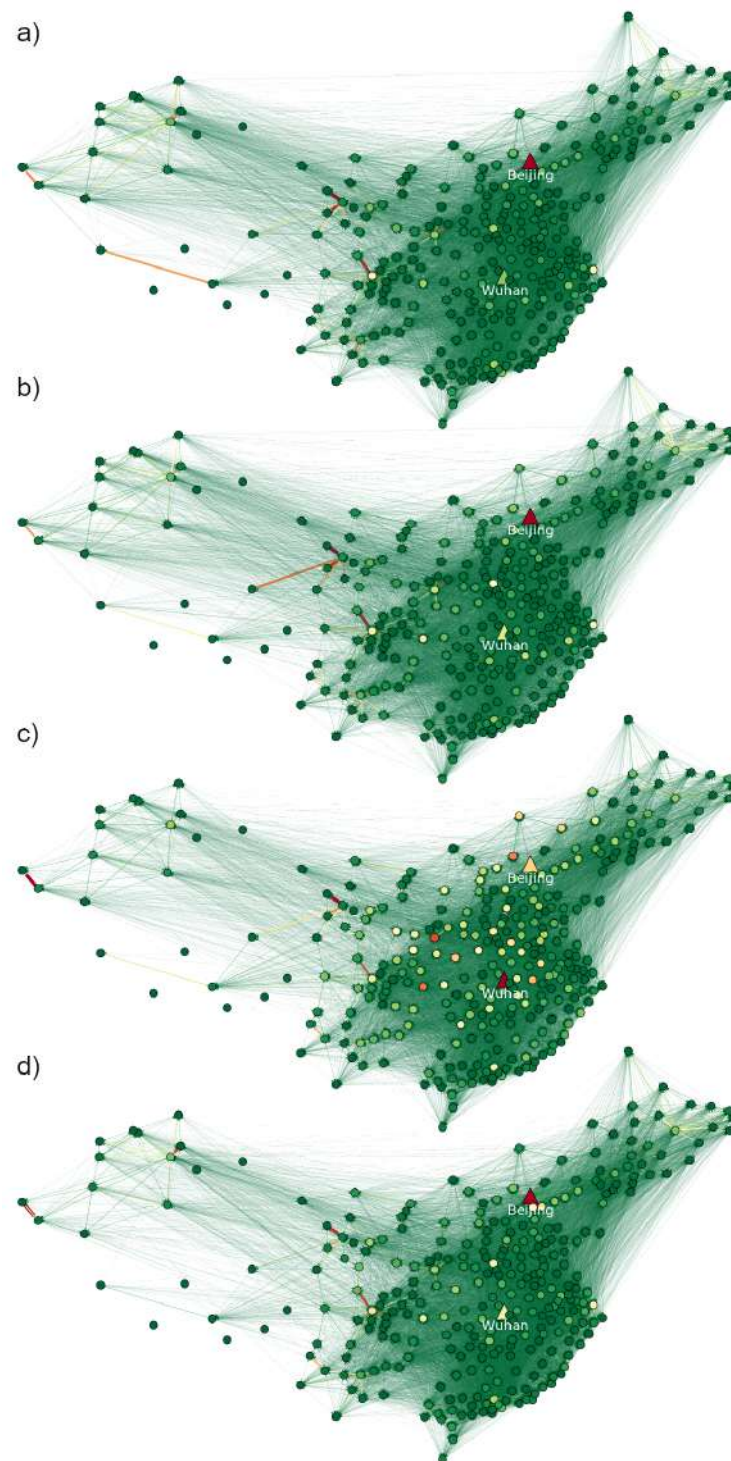


Figure 1. Mobility networks for: a) inflows and b) outflows of 2 January 2020; c) inflows and d) outflows of 2 February 2020. Points represent cities' centroids, and edges represent the flow of people between cities. Edge colors and widths are the flow intensities. Node colors are the weighted betweenness. Higher values are in red and lower in green.

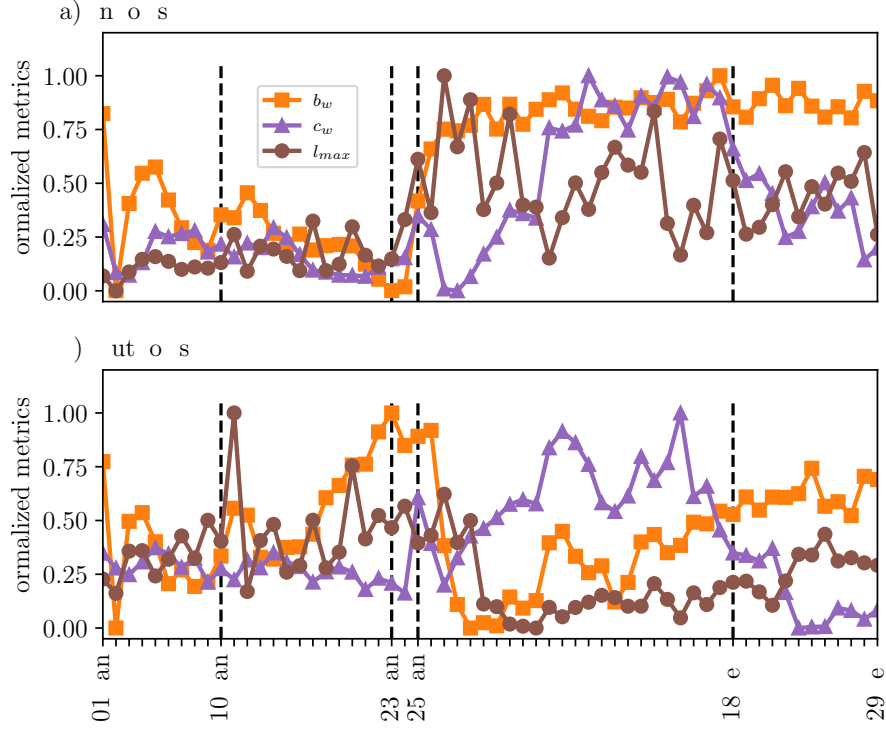


Figure 2. Time series of normalized network metrics for the a) inflow and b) outflow data. Their values lies within the following intervals a) $1,326.5 \leq b_w \leq 2,151.0$, $0.82 \leq c_w \leq 0.99$, $3.13 \leq l_{max} \leq 7.78$; b) $1,434.12 \leq b_w \leq 2,225.57$, $0.74 \leq c_w \leq 0.98$, $3.36 \leq l_{max} \leq 11.28$.

we use consists of two matrices for each day, one with the normalized inflow and another with the outflow between cities. Of all the people that arrive in a city (inflow), the data maps the contribution of all other cities in relative numbers (percentage). The same is valid for the people that leaves the city (outflow).

We show that the Wuhan travel ban on 23 January imposed changes to the level of importance of some central cities in the commuting patterns. While Beijing was the most important city in both the inflows and outflows, Wuhan and other cities became more relevant in the following days.

As future work, we intend to quantify correlations between the cities that are more central in the networks with the pandemic numbers.

5. Acknowledgement

The authors thank the Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Brazil (process: 441016/2020-0) and Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, CAPES, Brazil (process: 88887.506931/2020-00).

References

- Barabási, A.-L. and Pósfai, M. (2016). *Network science*. Cambridge University Press, Cambridge.
- China Data Lab (2020). Baidu Mobility Data.
- Cucinotta, D. and Vanelli, M. (2020). Who declares covid-19 a pandemic.
- Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track covid-19 in real time.
- Freitas, V. L. D. S., Konstantyner, T. C. R. D. O., Mendes, J. F., Sepetauskas, C. S. D. N., and Santos, L. B. L. (2020a). The correspondence between the structure of the terrestrial mobility network and the spreading of covid-19 in brazil. *Cadernos de Saude Publica*, 36:1–12.
- Freitas, V. L. D. S., Moreira, G., and Santos, L. B. L. (2020b). Robustness analysis in an inter-cities mobility network: modeling municipal, state and federal initiatives as failures and attacks toward sars-cov-2 containment. *PeerJ*, page e10287.
- Hu, T., Guan, W. W., Zhu, X., Shao, Y., Liu, L., Du, J., Liu, H., Zhou, H., Wang, J., She, B., Zhang, L., Li, Z., Wang, P., Tang, Y., Hou, R., Li, Y., Sha, D., Yang, Y., Lewis, B., Kakkar, D., and Bao, S. (2020). Building an open resources repository for covid-19 research. *Data and Information Management*, 4(3):130–147.
- Kraemer, M. U. G., Yang, C.-H., Gutierrez, B., Wu, C.-H., Klein, B., Pigott, D. M., null, du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., and Scarpino, S. V. (2020). The effect of human mobility and control measures on the covid-19 epidemic in china. *Science*, 368(6490):493–497.
- Lab, C. D. (2020). China COVID-19 Daily Cases with Basemap.
- Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., Ren, R., Leung, K. S., Lau, E. H., Wong, J. Y., et al. (2020). Early transmission dynamics in wuhan, china, of novel coronavirus–infected pneumonia. *New England Journal of Medicine*.
- Santos, L. B. L., Carvalho, L. M., Seron, W., Coelho, F. C., Macau, E. E. N., Quiles, M. G., and Monteiro, A. M. V. (2019). How do urban mobility (geo)graph’s topological properties fill a map? *Applied Network Science*, 4(1):91.
- Tan, S., Lai, S., Fang, F., Cao, Z., Sai, B., Song, B., Dai, B., Guo, S., Liu, C., Cai, M., Wang, T., Wang, M., Li, J., Chen, S., Qin, S., Floyd, J. R., Cao, Z., Tan, J., Sun, X., Zhou, T., Zhang, W., Tatem, A. J., Holme, P., Chen, X., and Lu, X. (2021). Mobility in China, 2020: a tale of four phases. *National Science Review*, 8(11).

Uso do Google Earth Engine como instrumento de identificação de anomalias climáticas no Brasil

Leonardo B. Prado¹ Gustavo E. Marcatti², Marconi A. Pereira³, José R. A. Júnior³
Filipe Santos³, Pedro A. A. Silva⁴, André Hirsch⁴, Daniel C. Pinheiro⁴, Andréa C. R. dos
Santos⁴, Fernando M. Pimenta⁴, Glauciana M. Ataíde², Renato V. O. Castro², Marcela
C. Ferreira², Weibert S. Pinto², Fernando A. Teixeira³, Gabriel Pereira¹

¹Departamento de Geociências - Universidade Federal de São João Del Rei (UFSJ) Caixa Postal: 35701-970 –
São João del Rei– MG – Brasil

²Departamento de Engenharia Florestal - Universidade Federal de São João Del Rei (UFSJ) Caixa Postal:
35701-970 – São João del Rei– MG – Brasil

³Departamento de Tecnologias em Eng. Civil, Computação e Humanidades -Universidade Federal de São João
Del Rei (UFSJ) Caixa Postal: 36490-972 – Ouro Branco – MG – Brasil

⁴Departamento de Ciências Agrárias – Universidade Federal de São João Del Rei (UFSJ) Caixa Postal
36307-352– Sete Lagoas – MG – Brasil

leonardoprado@gmail.com, gustavomarcatti@ufsj.edu.br, marconi@ufsj.edu.br, almeidajr.vr@gmail.com, filipesantos.lf@gmail.com, pedro@ufsj.edu.br, hirsch_andre@ufsj.edu.br, dcalbino@ufsj.edu.br, andreacr2010@gmail.com, fernando.m.pimenta@gmail.com, glauciana@ufsj.edu.br, renatocastro@ufsj.edu.br, mcferreira.edu@gmail.com, m.weibert.saturnino@gmail.com, teixeira@ufsj.edu.br, pereira@ufsj.edu.br

Abstract. *Due to the increase in extreme weather and climate events, this paper aims to identify the anomalies caused by such climatic variability and how they influence the agricultural sector, especially in the region known as MATOPIBA (Maranhão, Tocantins, Piauí e Bahia). For the climate analysis, reanalysis data from the European Center for Medium-Range Weather Forecasts (ECMWF) called Era5, with a spatial resolution of 11,132 meters, available in the Google Earth Engine database, were used. The results demonstrate that the water storage capacity in the soil, especially in the initial layers, is a variable of paramount importance for agricultural crops, as it includes climatic and edaphic aspects of the environment.*

Resumo. *Diante do aumento de eventos extremos de tempo e clima, o presente artigo tem como objetivo identificar as anomalias causadas por tais variabilidades climáticas e como as mesmas influenciam no setor agrário, principalmente na região conhecida como MATOPIBA (Maranhão, Tocantins, Piauí e Bahia). Para a análise climática utilizaram-se os dados de reanálise do European Centre for Medium-Range Weather Forecasts (ECMWF) denominado Era5, com resolução espacial de 11.132 metros, disponível na base de dados do Google Earth Engine. Os resultados encontrados demonstram que a capacidade de armazenamento de água no solo, principalmente nas camadas iniciais, é uma variável de suma importância para as culturas agrícolas, pois contempla aspectos climáticos e edáficos do ambiente.*

1. Introdução

Diante dos últimos relatórios desenvolvidos pelo *Intergovernmental panel on Climate change* (IPCC), principalmente a partir do relatório desenvolvido pelo IPCC AR6 (2021), não há dúvidas que há influência humana na variabilidade e na mudança climática. No Brasil, os produtores rurais vivenciam inúmeras dificuldades no campo, como o aumento da frequência de secas, chuvas concentradas, temperaturas elevadas e prolongamento das estiagens, que ameaçam a produtividade agrícola em diversas regiões brasileiras.

Além das alterações nas forçantes radiativas, variações climáticas de baixa frequência como, por exemplo, flutuações na temperatura do oceano Pacífico Equatorial relacionadas ao El Niño/ La Niña (ENOS) e Oscilação Decadal do Pacífico (ODP), alteram a distribuição da precipitação na América do Sul (CUNHA et al, 2011; REBOITA et al, 2010; REBOITA et al, 2012). O El Niño é um fenômeno atmosférico-oceânico caracterizado por um aquecimento anormal das águas superficiais no oceano Pacífico. Por outro lado, o La Niña caracteriza-se pelo esfriamento anormal nas águas superficiais do Oceano Pacífico (INPE, 2022). Neste contexto, compreender o efeito da variabilidade climática é de suma importância para reduzir os danos ocasionados na produtividade agrícola brasileira. Assim, o presente trabalho tem como objetivo principal utilizar o Google Earth Engine e sua base de dados para identificar e correlacionar as anomalias climáticas, especialmente a temperatura e a pluviosidade.

2. Metodologia

O ERA5-LAND é um produto de reanálise do *European Centre for Medium-Range Weather Forecasts* (ECMWF) que disponibiliza dados atmosféricos de 1981-presente. Estes dados combinam saídas de modelos numéricos de tempo e clima com observações *in situ* e base de dados de variáveis meteorológicas de todo o globo. Os dados do ERA5-LAND utilizados neste trabalho possuem tamanho de grade (*pixel*) de 11.132 metros e resolução temporal mensal. Ressalta-se que a base de dados disponível no Google Earth Engine (GEE) é um recorte temporal dos dados originais, desta forma, na plataforma estão disponíveis dados de 1982 até o presente. A partir da base de dados disponibilizada no GEE, extraiu-se as médias para cada mês do ano e as anomalias mensais de 2019 a 2021.

Os dados climáticos provenientes do ERA5-Land/ECMWF foram processados no Google Earth Engine. O GEE é uma plataforma para análise de dados em escala planetária e permite a utilização de diversas fontes de dados relacionadas à coleções de imagens provenientes de diversos satélites ambientais e meteorológicos, além de diversos produtos geoespaciais. A plataforma GEE possui dados de reanálises de modelos meteorológicos e dados climáticos obtidos por sensores orbitais. Neste trabalho a coleção do ERA5-Land ('ECMWF/ERA5_LAND/MONTHLY') foi recortada para as coordenadas geográficas 90° W/16° N; 30° W/57° S, correspondendo à América do Sul.

Para processar os dados no GEE utilizou-se o ambiente de desenvolvimento integrado (IDE) baseado na nuvem para o Earth Engine JavaScript Application Programming Interface (API). Neste, os dados mensais foram processados a partir de filtros temporais para estimar as médias mensais climatológicas, conforme equação 1:

$$A_{mes} = M - \Sigma M_i/n \quad (1)$$

em que a anomalia para respectivo mês (A_{mes}) é obtida pelo valor do mês correspondente ($M_{i,12}$) subtraído a média climatológica para o referido mês. No GEE a média histórica mensal é obtida pela biblioteca *ee.Filter.calendarRange*.

4. Resultados e discussão

A Figura 1 representa a média de temperatura do ar a 2 metros, precipitação e umidade do solo nas camadas superficiais durante o período de (1982-2021). Durante o período analisado, as maiores temperaturas médias estão concentradas na região equatorial da América do Sul. Em relação à precipitação, percebe-se um volume maior na região amazônica e na região Sul do Brasil, nestas, os volumes anuais excedem 1.750 mm. Ainda, devido principalmente à Zona de Convergência do Atlântico Sul (ZCAS) é possível observar uma faixa no sentido noroeste-sudeste próxima ao Estado de São Paulo. Em relação a umidade no solo, nas camadas de 0 a 7 cm e 7 a 28 cm, as regiões brasileiras que apresentaram as menores médias estão localizadas na Caatinga e sudeste do Brasil.

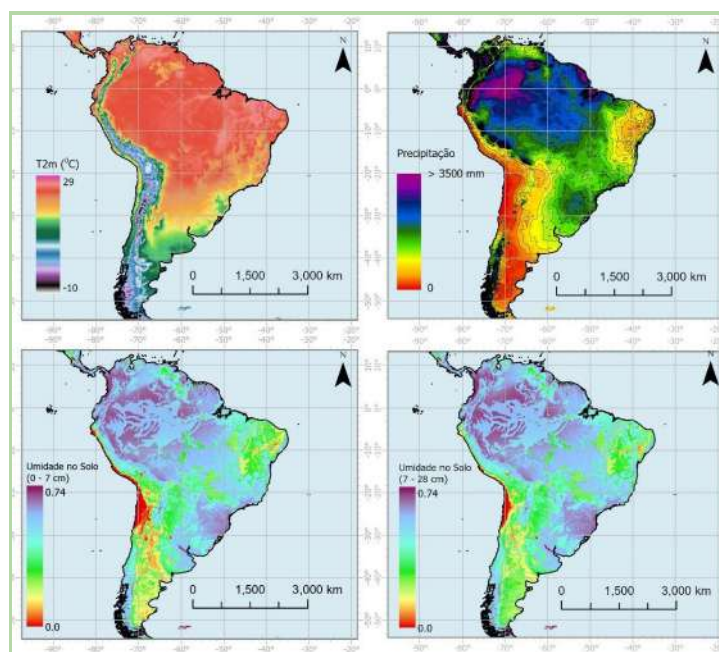


Figura 1: Climatologia do Brasil no período de (1982-2021).

As Figuras 2 e 3 representam as anomalias de precipitação e temperatura do ar a 2 metros, de setembro de 2019 a dezembro de 2020, para a América do Sul. Ao analisar os mapas de anomalia da precipitação foi possível constatar que o ano de 2020 foi um ano mais seco que o normal, devido aos menores valores de precipitação. A escassez hídrica de grande parte destas áreas pode ser explicada pela atuação do fenômeno La Niña de intensidade

moderada. Ressalta-se que anomalias na precipitação foram acentuadas nos meses de setembro, outubro, novembro e dezembro, períodos importantes para o plantio de commodities no Brasil.

Tais dados estão de acordo com os boletins agroclimatológicos publicados pelo Inmet (2021) que relacionam eventos importantes de seca no Brasil e atrasos nas datas de plantio devido à escassez hídrica no ano de 2020. Regiões como Sudeste, Sul e Centro-Oeste do país, bem como, novas fronteiras agrícolas como MATOPIBA (Maranhão, Tocantins, Piauí e Bahia) e SEALBA (Sergipe, Alagoas e Bahia), foram as regiões que mais sofreram com a seca.

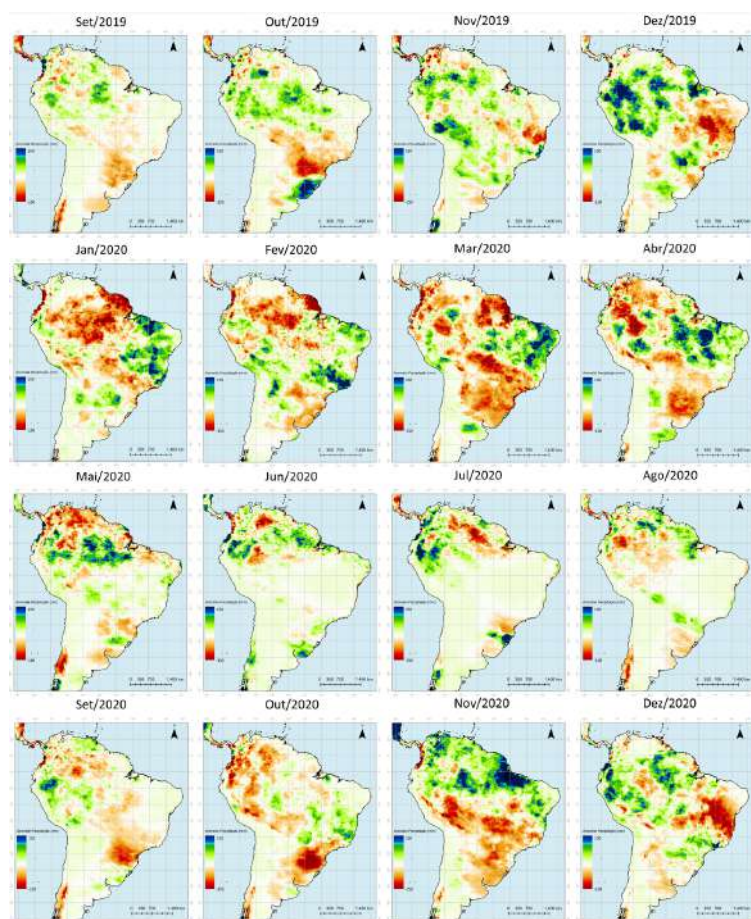


Figura 2: Anomalia da Precipitação de Setembro de 2019 à Dezembro de 2020.

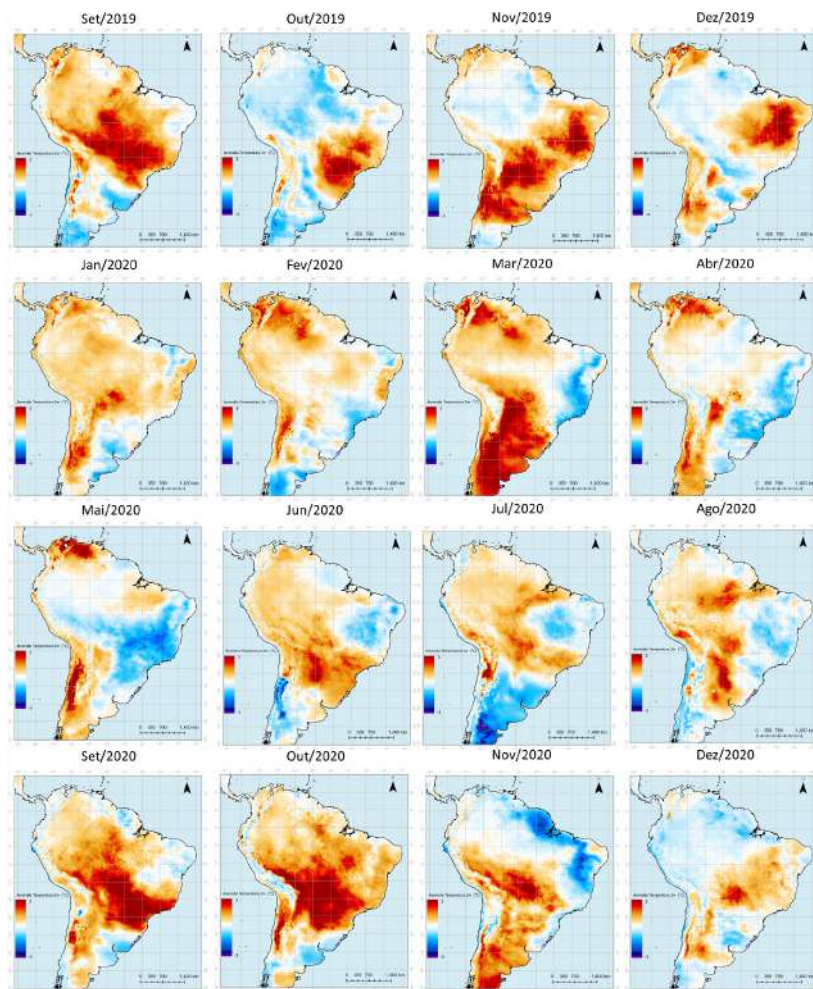


Figura 3: Anomalia da temperatura de Setembro de 2019 à Dezembro de 2020.

Foi possível observar através dos mapas produzidos que a a escassez de água no solo representado pela figura 2, durante o ano de 2020 possuiu uma flutuabilidade, ao observar principalmente as novas fronteiras agrícolas (MATOPIBA e SEALBA), se observou que entre Janeiro e Abril de 2020 a anomalia de precipitação tornou-se positiva, isto é, o volume de chuvas da região estava acima da média histórica, enquanto nos meses posteriores a anomalia positiva reduziu, se aproximando da neutralidade mas mantendo-se positiva até Agosto do mesmo ano, do mês de Setembro até Dezembro é possível constatar que ocorre uma instabilidade nas anomalias nas novas fronteiras agrícolas, chamando a atenção para o mês de dezembro com o predomínio de escassez hídrica.

O mapa de anomalias de temperatura nos apresenta que em todos os meses a anomalia positiva de temperatura atingiu o seu máximo, diferenciando-se pela espacialidade do fenômeno pelas regiões do Brasil, as áreas do MATOPIBA e SEALBA durante os últimos quatro meses de 2019 foram fortemente influenciadas pelas altas temperaturas, mas não sendo as únicas a serem atingidas pelo fenômeno. Durante os meses de Maio de 2020 e Agosto de 2020 foi possível constatar que as temperaturas no nordeste mantiveram-se abaixo

do normal ou mesmo apresentaram neutralidade. Fato que não perdurou para os meses seguintes (Setembro à Dezembro de 2020) destacando-se o mes de Outubro para uma anomalia térmica positiva dominante por todo o território brasileiro.

5. Considerações Finais

Considerando que a umidade do solo depende não somente da precipitação, mas também da capacidade de armazenamento de água do solo, os mapas de anomalia do volume de água nas camadas superficiais do solo podem ser mais importantes que o de precipitação, considerando que tal variável contempla aspectos climáticos e edáficos do ambiente. Desta forma, os elementos climáticos permitem uma análise sistemática dos processos da interação entre terra-atmosfera e a produtividade agrícola, principalmente nas novas áreas de fronteira agrícola como MATOPIBA e SEALBA.

Agradecimentos

Os autores agradecem ao Ministério da Agricultura, Pecuária e Abastecimento - MAPA (processo 23122.040239/2021-74) e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais - FAPEMIG (projeto APQ-00718-21) pelo suporte financeiro e à Universidade Federal de São João del Rei - UFSJ pela infraestrutura e fomento a pesquisa.

Referências bibliográficas

INMET, 2021. Boletins Agroclimatológicos [WWW Document]. URL <<https://portal.inmet.gov.br/boletinsagro>>. acesso em 6/6/2022.

IPCC AR6 (Intergovernmental Panel on Climate Change). Summary for Policymakers. In: MASSON-DELMOTTE, V., P. ZHAI, A. PIRANI, S. L. CONNORS, C. PÉAN, S. BERGER, N. CAUD, Y. CHEN, L. GOLDFARB, M. I. GOMIS, M. HUANG, K. LEITZELL, E. LONNOY, J.B.R. MATTHEWS, T. K. MAYCOCK, T. WATERFIELD, O. YELEKÇI, R. YU AND B. ZHOU (Eds). Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press, 2021, 41 pp.

REBOITA, M. S, GAN, M. A, ROCHA, R. P. D. & AMBRIZZI, T. Regimes de precipitação na América do Sul: uma revisão bibliográfica. **Revista brasileira de meteorologia**, 25, 185-204, 2010.

REBOITA M. S., KRUSCHE, N., AMBRIZZI, T., & ROCHA, R. P. D. Entendendo o Tempo e o Clima na América do Sul, 2012.

NULL, J. El Niño and La Niña Years and Intensities, 2022. Disponível em <<https://ggweather.com/enso/oni.htm>> . Acesso em 05/09/2022

Uma Metodologia para Apoio à Tomada de Decisão em Cenários de Desastres Ambientais Utilizando Dados Públicos

Carlos Henrique Tavares Brumatti, Mariana Albuquerque Reynaud Schaefer,
Jugurta Lisboa-Filho, Julio C. S. Reis

¹ Departamento de Informática – Universidade Federal de Viçosa (UFV)
Viçosa – MG – Brasil

{carlos.h.tavares,mariana.schaefer,jugurta,jreis}@ufv.br

Abstract. *Environmental disasters, such as dam failures, cause impacts that go beyond the area of occurrence. From the region of origin to its arrival at sea, the sediments can cause environmental and economic impacts. Searching for ways to help in the recovery of these degraded areas, this work proposes the development of a methodology in that users utilize a Knowledge Discovery in Database Process (KDD) to group cities affected by environmental disasters, thus presenting the generation of knowledge groups with the same characteristics. In addition, reflecting all aspects of the proposed methodology, the development of an application that can be made available to the end user is planned. We hope that the proposed framework can be useful to support expert decision-making in this context.*

Resumo. *Desastres ambientais, como o rompimento de barragens, causam impactos que vão muito além da área de ocorrência. Da região de origem até a sua chegada ao mar, os resíduos podem causar tanto impactos ambientais quanto econômicos. Buscando formas de auxiliar na recuperação dessas áreas degradadas, neste trabalho, é proposto o desenvolvimento de uma metodologia utilizando o processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Database – KDD) a fim de agrupar cidades afetadas por desastres ambientais, gerando assim grupos de cidades com características semelhantes. Além disso, refletindo toda a metodologia proposta, planeja-se o desenvolvimento de uma aplicação que possa ser disponibilizada para o usuário final. Esperamos que o arcabouço proposto possa ser útil para suportar a tomada de decisões de especialistas neste contexto.*

1. Introdução

O desastre ambiental causado pelo rompimento da barragem de Fundão, em Mariana (MG), no dia 05 de novembro de 2015, evidenciou a íntima relação existente entre o meio biótico e o meio antrópico. Segundo Godoy e Dias [Godoy and Dias 2021], o impacto causado pelo referido desastre provocou sérias consequências não somente na região de ocorrência, mas em todo o percurso dos rejeitos até a sua chegada ao mar. Ademais, estas consequências não se restringem somente às questões ambientais, mas também socioeconômicas, políticas e humanas.

Nesse contexto, um dos problemas identificados, por exemplo, está relacionado à alocação de recursos financeiros, de modo adequado e eficiente, para a recuperação das

áreas afetadas. Segundo Barbosa et al [Barbosa et al. 2015], após o desastre, estima-se que as prefeituras das áreas envolvidas terão que gastar cerca de R\$150 milhões, além de que há uma proposta da criação de um fundo de US\$20 bilhões, ao longo de 10 anos, pelas empresas envolvidas.

Diante deste cenário, soluções computacionais podem ser úteis para apoiar especialistas no processo de tomada de decisão relacionado a desastres ambientais. Uma das maneiras existentes para isso é através do processo de Descoberta de Conhecimento em Bases de Dados (do inglês, *Knowledge Discovery in Database - KDD*), principalmente na etapa de Mineração de Dados. Segundo Camilo e Silva [Camilo and Silva 2009], tais técnicas possibilitam a resolução de tarefas como Descrição, Classificação, Estimação, Predição, Agrupamento e Associação. Fundamentado em KDD, neste trabalho é proposta uma metodologia que permite agrupar cidades atingidas por desastres ambientais utilizando dados públicos. Acreditamos que o arcabouço proposto, juntamente com o desenvolvimento de uma aplicação que o abarque, possa ser útil para suportar a tomada de decisões de especialistas neste contexto, como por exemplo, no processo de decisão de cidades (ou áreas) atingidas por desastres ambientais e que, em detrimento de alguma característica (em comum), cuja recuperação deve ser priorizada.

O trabalho foi organizado conforme detalhado a seguir. A Seção 2 apresenta a fundamentação teórica. Na Seção 3 é descrita a metodologia proposta. Depois, na Seção 4 apresenta os resultados e discussões preliminares. Por fim, na Seção 5 são apresentadas as conclusões e direções para trabalhos futuros.

2. Fundamentação Teórica

Das várias tarefas que a Mineração de Dados se propõe a resolver, destaca-se o Agrupamento. Segundo Jain et al. [Jain et al. 1999], considerando um conjunto de dados, as técnicas voltadas para tal tarefa buscam gerar agrupamentos (ou *clusters*) baseados na similaridade dos elementos contidos em um mesmo grupo. Essa similaridade é um critério que define o quanto dois ou mais elementos são semelhantes e, conseqüentemente, devem pertencer a um mesmo conjunto gerado.

Considerando os algoritmos existentes, os mais tradicionais, segundo Cassiano [Cassiano 2014], podem ser classificados como ilustrado na Figura 1. No primeiro nível da classificação, os algoritmos se dividem em relação a abordagem: hierárquica ou particional. A abordagem hierárquica, também segundo Cassiano [Cassiano 2014], se caracteriza por manter o par de dados mais próximo juntos. Já a abordagem particional divide a base de dados em k grupos, sendo o número k um valor informado pelo usuário, ainda conforme Cassiano [Cassiano 2014]. No segundo nível, tem-se a subdivisão da abordagem de acordo com a forma de medição da similaridade. Das várias técnicas de Agrupamento existentes, optou-se pelo *K-Means* como base para a metodologia aqui proposta, conforme detalhado a seguir¹.

2.1. *K-Means*

O Algoritmo *K-Means* é um algoritmo clássico e bastante explorado na literatura e foi proposto em 1967 [MacQueen 1967]. Por ser mais simples de implementar e de baixa

¹Neste contexto, é válido mencionar que embora o *K-Means* tenha sido selecionado para aplicação nesta versão preliminar da metodologia proposta, ela é genérica para abarcar outras técnicas que poderão ser investigadas futuramente

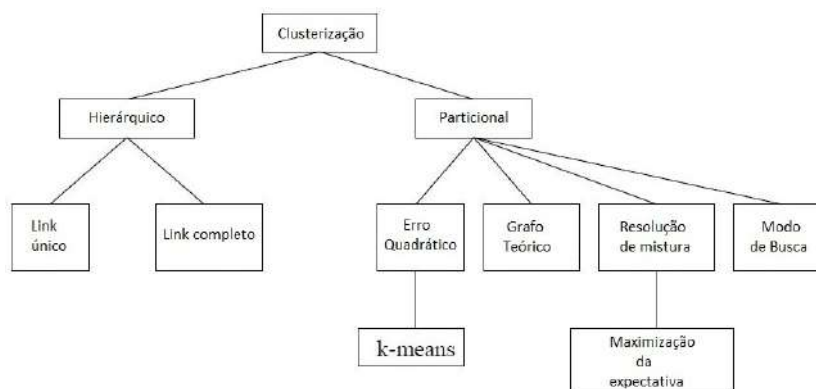


Figura 1. Classificação dos algoritmos de clusterização. Fonte: [Jain et al. 1999]

complexidade, é comumente empregado para a geração de grupos. Seu pseudocódigo é apresentado abaixo. Este algoritmo se baseia na tentativa de minimizar o erro quadrático calculado associado a distância entre cada elemento e o centróide do seu respectivo grupo, como medida de similaridade.

Algorithm 1 Pseudocódigo do Algoritmo *K-Means*

- 1 Especifique o número K de agrupamentos para gerar
 - 2 Inicialize os K centróides de forma randômica
 - 3 **while** A posição do centróide não mudar **do**
 - 4 Atribua cada ponto ao centróide mais próximo
 - 5 Calcule o novo valor de centróide (média) de cada agrupamento
- end**
-

3. Metodologia Proposta

Uma visão geral da metodologia proposta neste trabalho é apresentada na Figura 2. Nela, é realizada também a associação das etapas de trabalho proposta com cada uma das 4 etapas do KDD apresentadas por Miller e Han [Miller and Han 2009].

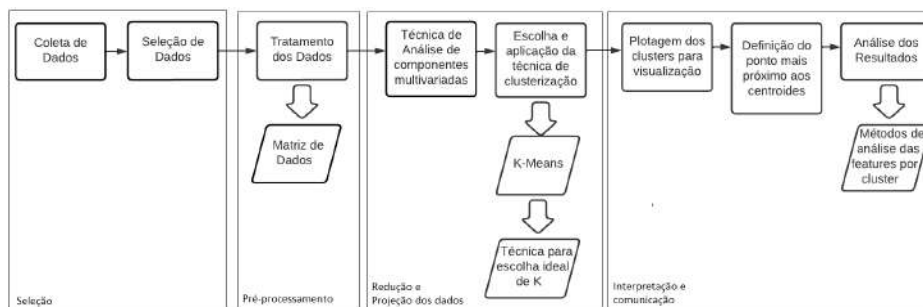


Figura 2. Visão geral da metodologia proposta.

A Etapa 1, denominada de Seleção, é onde ocorre a coleta e a seleção dos dados de interesse. Para esse processo, optou-se pelo uso dos dados do Cadastro Ambiental

Rural (CAR)² e do Instituto Brasileiro de Geografia e Estatística (IBGE)³. Verificou-se que os dados do CAR são pouco explorados em trabalhos desse tipo, enquanto o IBGE é uma fonte importante devido ao grande volume de informações presentes. Além disso, tem-se que os dados do CAR fornecem uma boa descrição da zona rural das cidades, uma vez que contém informações referentes ao número de Imóveis Rurais, de Áreas de Proteção Permanente (APP), dentre outros. Já os dados do IBGE são interessantes por descreverem informações socioeconômicas importantes, como Densidade Demográfica e Taxa de Escolarização, por exemplo. Após a coleta dos dados, é realizada a seleção dos dados mais interessantes para o processo. Nesse momento, é de interesse e/ou desejável a presença de um usuário especialista para verificar quais dados possuem de fato impacto na regra de negócio a ser estudada.

Na próxima etapa, denominada de Pré-Processamento (2), os dados, selecionados pelo usuário especialista, passam pelo processo de tratamento. Assim, dados de diversas fontes e formatos são agora ajustados conforme a necessidade do algoritmo a ser utilizado. Por exemplo, para a metodologia proposta, os dados categóricos são convertidos em dados numéricos e os dados em ponto flutuante, dependendo do formato de entrada, precisam ter sua representação alterada, facilitando assim o processo.

Na sequência, na Etapa 3, ocorre a Redução e Projeção dos dados. Como para a metodologia aqui descrita é usado o algoritmo *K-Means* para a geração dos agrupamentos, agora é aplicada uma técnica de análise de componentes multivariadas, fazendo com que a matriz de dados, gerada na etapa anterior, seja convertida em uma matriz dimensional. Para isso, utilizou-se o *Principal Component Analysis* (PCA) [Bro and Smilde 2014]. Como o *K-Means* é um algoritmo de clusterização particional, o valor do número de agrupamentos k foi determinado usando o *Silhouette Score* [Rousseeuw 1987], que de maneira geral captura a consistência dentro de um determinado agrupamento de dados. É importante mencionar que, embora tenham sido aplicados algoritmos e/ou técnicas específicas baseadas em critérios pré-definidos, a metodologia é generalizável. Em outras palavras, caso seja de interesse do usuário, os algoritmos e/ou técnicas explorados em alguns componentes podem ser alterados.

Por fim, na Etapa 4, logo após a geração de todos os agrupamentos e a determinação de seus centroides, ocorre a visualização dos dados. Nesta etapa é definido também o ponto intra-conjunto mais próximo do centroide determinado. Assim, esse ponto é o elemento mais indicado para ser um representante do agrupamento, ou seja, um elemento de referência que possuirá as propriedades mais características do conjunto a que pertence, já que é o ponto de maior similaridade do conjunto.

4. Resultados e Discussões Preliminares

Refletindo a metodologia anteriormente descrita, iniciou-se a implementação de um protótipo de um Sistema de Inteligência Geográfica, conforme a arquitetura ilustrada na Figura 3. De forma geral, a ideia geral desta plataforma proposta é permitir que o usuário (especialista) realize o *upload* dos seus dados (de interesse) para aplicação da metodologia descrita na seção anterior.

²<https://www.car.gov.br/>

³<https://www.ibge.gov.br/>

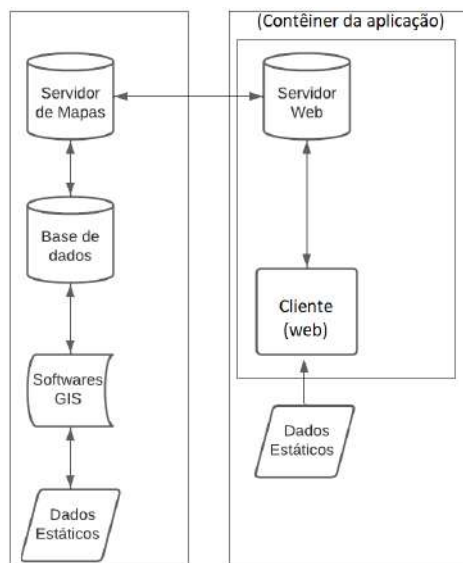


Figura 3. Arquitetura geral da aplicação proposta.

Para a construção dessa plataforma *Web*, utilizou-se tecnologias como *HyperText Markup Language* (HTML), *Cascading Style Sheets* (CSS) e Javascript. Já para o *backend*, usou-se o *Python*, uma vez que os principais algoritmos para uso em Mineração de Dados já se encontram implementados nessa linguagem. Além disso, o sistema como um todo é conectado a um servidor de mapas, provendo assim para a aplicação dados nos formatos *Web Map Service* (WMS) e *Web Feature Service* (WFS), conforme especificações da *Open Geospatial Consortium* (OGC)⁴. A Figura 4 ilustra um resultado preliminar da aplicação da metodologia proposta, baseada inicialmente no algoritmo *K-Means*, ao conjunto de dados já obtido para a Bacia Hidrográfica do Rio Paraopeba (BHRP) (ver detalhes da Seção 3). Essa região foi escolhida devido ser uma das áreas afetadas pelo rompimento da barragem da Mina Córrego do Feijão, em Brumadinho (MG), no ano de 2019. Na Figura, cada um dos alfinetes (ou pontos) representa uma das cidades que compõem a referida bacia. Além disso, cada cor representa um dos agrupamentos distintos obtidos. Para esse exemplo, foram obtidos três agrupamentos, representados pelas cores vermelho, verde e amarelo. A partir disso, o usuário pode identificar de forma visual quais as cidades que são mais semelhantes e, em seguida, buscar extrair informações estatísticas a cerca dos dados relacionados, verificando assim quais as áreas que mais necessitam dos investimentos.

5. Conclusão

Este artigo apresentou uma metodologia genérica que busca auxiliar os usuários responsáveis pelo processo de tomada de decisão, seja ele pertencente a um órgão público ou empresa privada, durante a ocorrência de desastres ambientais que causam impacto direto nas cidades próximas. Espera-se que o trabalho proposto possa ser útil para apoiar decisões relacionadas à esses desastres ambientais, provendo aos especialistas informações que possam suportar decisões neste contexto. Como trabalhos futuros, planeja-se disponibilizar a ferramenta para uso, realizando a coleta de *feedbacks* dos usuários e imple-

⁴<https://www.ogc.org/>



Figura 4. Plotagem dos agrupamentos obtidos como exemplo para a BHRP.

mentando novas funcionalidades. Além disso, planeja-se também testar outros algoritmos de Agrupamento para comparação dos resultados obtidos.

Agradecimentos. Trabalho parcialmente financiado por CNPq, CAPES e FAPEMIG.

Referências

- Barbosa, F. A. R., Maia-Barbosa, P. M., Nascimento, A. M. A., Rietzler, A. C., Franco, M. W., Paes, T. A., Reis, M., Moura, K. A. F., Dias, M. F., de Paula Ávila, M., et al. (2015). O desastre de mariana e suas consequências sociais, econômicas, políticas e ambientais: porque evoluir da abordagem de gestão dos recursos naturais para governança dos recursos naturais? *Arquivos do Museu de História Natural e Jardim Botânico da UFMG*, 24(1-2).
- Bro, R. and Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, 6(9):2812–2831.
- Camilo, C. O. and Silva, J. C. d. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFG)*, 1(1):1–29.
- Cassiano, K. M. (2014). Análise de séries temporais usando análise espectral singular (ssa) e clusterização de suas componentes baseada em densidade. *Pontifícia Universidade Católica do Rio de Janeiro*.
- Godoy, S. M. and Dias, M. B. (2021). O desastre ambiental de mariana e o papel da fundação renova na reparação dos danos. *Direito e Desenvolvimento*, 12(1):37–48.
- Jain, A. K., Murty, M. N., and Flynn, P. J. (1999). Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323.
- MacQueen, J. (1967). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability*, pages 281–297.
- Miller, H. J. and Han, J. (2009). *Geographic data mining and knowledge discovery*. CRC press.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.

Analysis of the watersheds of National Interconnected System (SIN) using Complex Network

Angélica N. Caseri¹, Francisco Aparecido Rodrigues¹

¹Instituto de Matemática e Ciência da Computação – Universidade de São Paulo (USP)

angelica.caseri@gmail.com, francisco@icmc.usp.br

Abstract. *Os recursos hídricos no Brasil são uma das maiores fontes de geração de energia do país. O Sistema Interligado Nacional (SIN) é constituído por diversas bacias que são interligadas. Acompanhar e monitorar estes reservatórios de água é de grande importância para o Brasil. Ao longo dos anos é possível identificar alterações no sistema hídrico devido a fatores climáticos, como as mudanças climáticas. Este trabalho tem como objetivo identificar a conexão entre as bacias hidrográficas que constituem o SIN, a correlação entre as suas series históricas de Energia Natural Afluente (ENA) utilizando técnicas de redes complexas. Por meio deste trabalho é possível identificar as bacias que possuem maior número conexões e que são de elevada importância no sistema, as suas características influenciam nos fatores de outras bacias.*

Resumo. *Water resources in Brazil are one of the largest sources of energy generation in the country. The National Interconnected System (SIN) consists of several watersheds that are interconnected. Tracking and monitoring these water reservoirs is of great importance for Brazil. Over the years it is possible to identify changes in the water system due to climatic factors, such as climate change. This work aims to identify the connection between the hydrographic watersheds that constitute the SIN, the correlation between their historical series of Affluent Natural Energy (ENA) using complex network techniques. Through this work it is possible to identify the watersheds that higher degree connections and that are of high importance in the system, their characteristics influence the factors of other watersheds.*

1. Introduction

The climatic conditions directly affect renewable energies and, consequently, the production of electricity. The generation of hydropower depends heavily on the volume of water in the reservoirs to feed the hydroelectric turbines (Queiroz et al. 2016). In addition, this volume of water depends on precipitation (climate variable), which is often represented in rainfall-runoff models (Collischonn et al., 2007; Caseri et al., 2022). Climate change induced by global warming can lead to an increase in the frequency and intensity of extreme weather events. The renewable energy sector is subject to the adverse effects of climate change. Changes in rainfall patterns can affect energy production from different sources (Lelis et al., 2011). Hydroelectric power dominates energy sources in Brazil, contributing 62% of its total installed electricity generation capacity at the end of 2020. Brazil ended 2020 with a power generation capacity of 174,413 MW, of which 58.6% generated by hydro, 25.2% thermal, 9.64% wind, 3.1% small hydro, 1.87% fotovoltaic, 1.13% nuclear and 0.47% mini hydro (HydroReview, 2021).

In order to understand climate phenomena and anticipate them, methods such as complex networks and statistics analysis are used in several research works (Donges et al., 2009; Rodrigues and Costa, 2010). Complex networks are a set of nodes with connections to each other. They are characterized by a structure capable of statistically investigating the topology of interrelationships (Donges et al., 2009). These nodes can represent, for example, locations, physical objects. In turn, connections can have different interpretations, such as mathematical relationships, conceptual affiliations (Rodrigues and Costa, 2010).

This work aims to identify the most important watersheds within the Brazil Hydrological System considering Pearson correlation and complex networks. Thus, identifying the watersheds and locations that may suffer the highest impact between the watersheds. The main differential of this work is to combine statistical techniques, complex networks and combine different data to identify the connection on watershed systems.

2. Study area and data

The National Interconnected System (SIN) is divided into 4 segments called subsystems:

- Southeast/Midwest (SE/CO): covers Rondônia and Acre in addition to the Southeast and Midwest regions of Brazil;
- South (S): covers the southern region of Brazil;
- Northeast (NE): covers the entire northeast region, with the exception of Maranhão;
- North (N): covers the north of Brazil, encompassing the states of Amapá, Amazonas, Maranhão, Pará and Tocantins.

ENA is Affluent Natural Energy, amount of water from a hydroelectric plant that can be used to generate energy. It represents the energy that can be produced with the flow of water from a given river to a reservoir of a hydroelectric plant. ENA Storage considers the natural flows discounted from the flows poured into the reservoirs. ENA can be measured in MWm (averages) and as a percentage of the historical average.

In this analysis, the daily ENA history for all watersheds was considered of 2012 – 2020 to identify possible changes in characteristics over the years and the correlation of the time series from different watersheds. In this study we consider the following watersheds: Amazonas, Araguari, Parnaíba, São Francisco, Paraguaçu, Jequitinhonha, Mucuri, Doce, Itabapoana, Paraíba do Sul, Capivari, Itajaí, Jacuí, Uruguai, Paraná, Paraguai, Tocantins, Grande, Iguaçu, Paranapanema and Tiête.

3. Results

To understand the behavior of ENA data throughout the year, time series analyzes were used to compare the behavior of ENA over the months and years from complex networks (graphs), and correlation analysis of the time series. Graphs are structures that represent a set of items and the relationship between them. A graph $G(V, E)$ is composed of vertices that connect through edges. NetworkX is a Python package used for handling

graphs and networks. In this study we used the NetworkX to generate the graphs to understand the connectivity between the watersheds to identify those with the highest connection with the other. The Figure 1 presents the complex network from the principal watersheds from SIN considering their spatial connection, that is, the watersheds that are neighbors.

From the analysis of the complex network, it is noted that the hydrographic watersheds that have the highest number of connections are the São Francisco and Paraná watersheds, with connection to 8 watersheds. The Tocantins watershed also plays an important role, it has numerous connections, there are 5 connections in total. It is also verified that the watersheds Capivari, Araguaí and Paraguaçu have only one connection, they are isolated from the rest of the network.

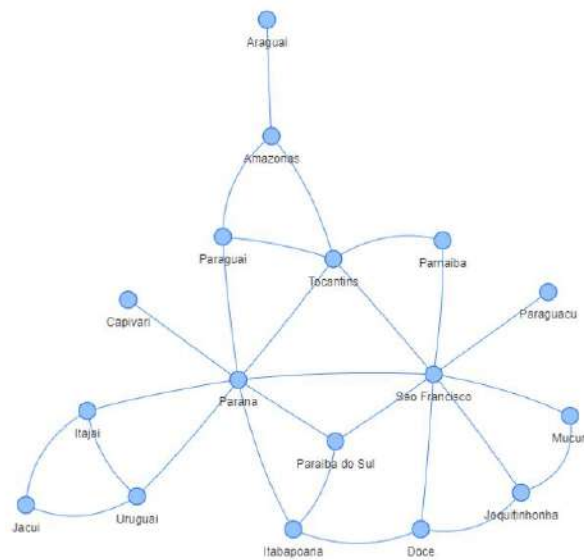


Figure 1. Complex network of watersheds.

With the aim of understanding the importance between the watersheds and the correlation between their time series, in this study the correlation matrix of the ENA time series was calculated. Additionally, graphs were built which the edge weights are based on the correlation value between the time series watersheds (strong color greater correlation, blue negative correlation and green positive correlation). Figures 2 to 4 present the results of these analyses.

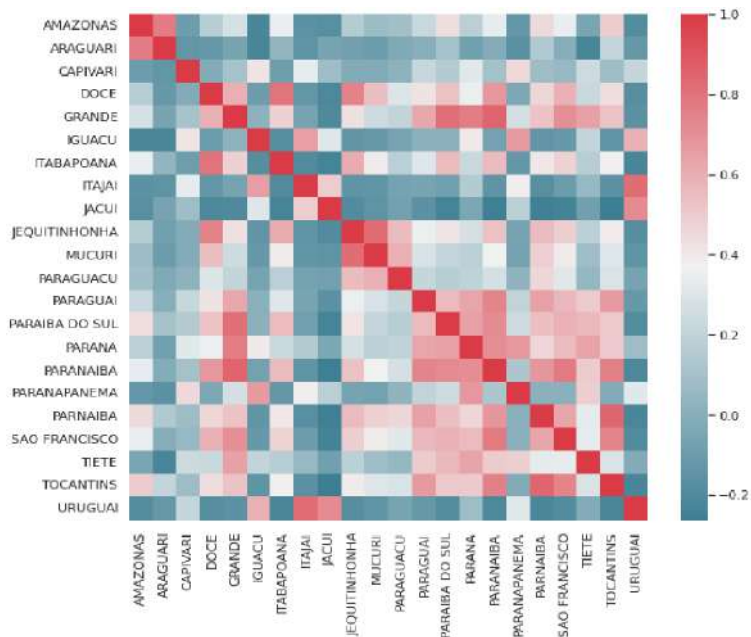


Figure 2. correlation matrix of watershed time series.

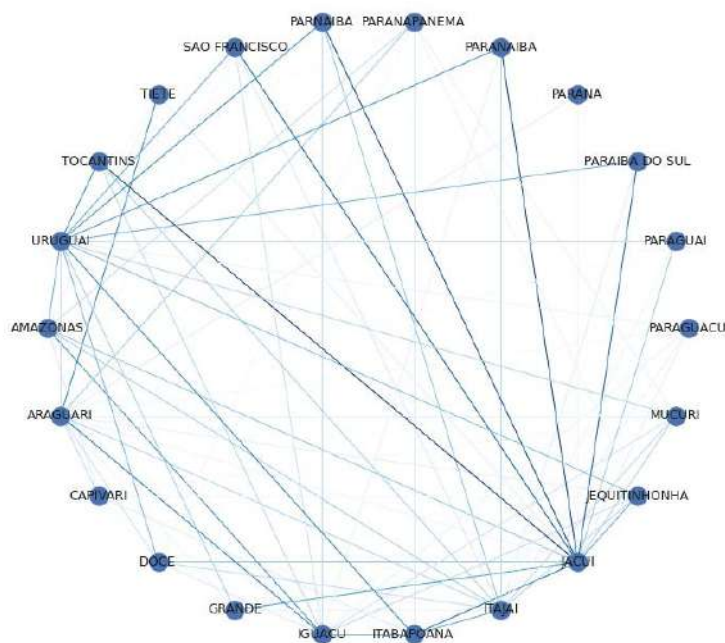


Figure 3. Complex network of negative correlations of watershed time series (the stronger the color, the greater the correlation).

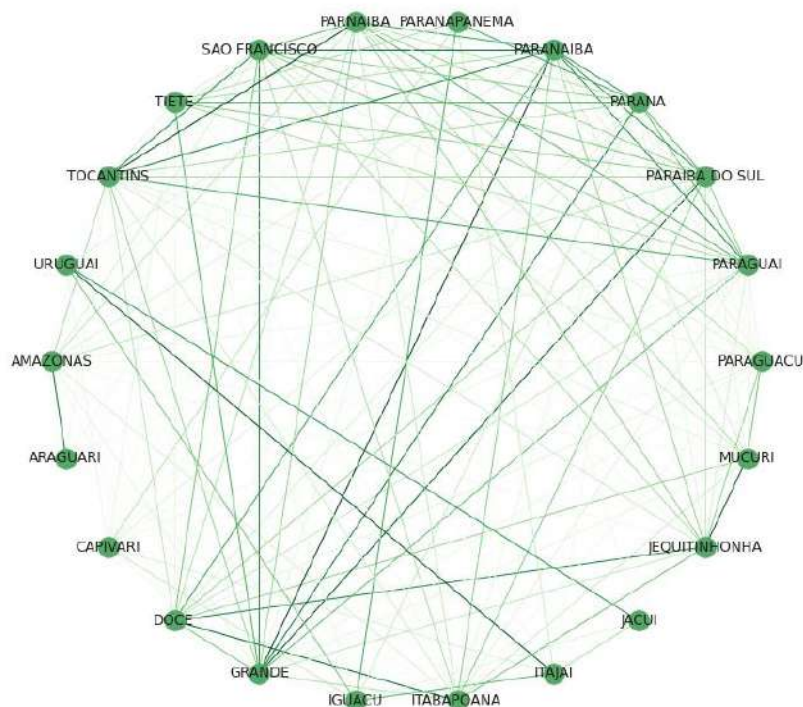


Figure 4. Complex network of positive correlations of watershed time series (the stronger the color, the greater the correlation).

Through the correlation matrix and the graphs built through weights based on the correlation matrix, the Jacui watershed has a high negative correlation with several watersheds, such as Parnaíba, São Francisco, Paranapanema. As well as the Tocantins, Paraguai, São Francisco, Paraná and Paranaíba watersheds, they have a high positive correlation with several watersheds. It is also verified that the majority of the watersheds have correlation with each other, with few cases that the correlation is close to zero, as is the case of the São Francisco watershed with Mucuri, for example.

Comparing the graphs based on the weights of the correlations between the watersheds and the complex network that considers the proximity between the watersheds, it appears that in some cases nearby watersheds have low correlation, such as for example Amazonas and Paraguay, Parana and Capivari. However in other cases there is a high correlation, such as São Francisco and Doce. Equally, it is also possible to notice watersheds that have high correlation but that are not neighboring Jacui and Itabapoana. In this way, it is observed that other factors have influence on the correlation between watersheds, not just the proximity between them. One can exemplify climatic factors and the hydrological system of the watersheds.

4. Conclusion

In general, through the analyzes carried out, it is possible to observe that the hydrographic watersheds have a high connection with each other, some of which are essential for the functioning of a large part of the hydrological system in Brazil. The São Francisco and Paraná watersheds are the ones that have the highest connection with the other watersheds. However, from the analysis of graphs calculated through the correlation between the time series of the watersheds, it is noted that the Jacuí, Tocantins, Paraguai, São Francisco, Paraná and Paranaíba watersheds have high importance for the system, as they have high correlation with many hydrological watersheds, their time series are highly correlated with the other watersheds.

As next steps of this work, descriptive analysis of other variables, such as EARM (Maximum Stored Energy) and rainfall data in order to generate projections of the reservoirs through the use of deep learning algorithms.

References

- CASERI, N.A.; SANTOS, L.S., STEPHANY, S. “A Convolutional Recurrent Neural Network for Strong Convective Rainfall Nowcasting Using Weather Radar Data in Southeastern Brazil”, *Artificial Intelligence in Geosciences*. v.3, 2022.
- COLLISCHONN, W., Allasia, D.G., SILVA, B.C., TUCCI, C.E.M. “The MGB-IPH model for large-scale rainfall-runoff modeling”, *Hydrological Sci. J.* 2007.
- DE QUEIROZ, A.R., MARANGOM LIMA, L. M. MARANGIM LIMA, J.W. DA SILVA, B.C., SCIANNI, L.A. “Climate change impacts in the energy supply of the Brazilian hydro-dominant power system”, *Renewable Energy*. 99, 2016.
- DONGES, J.F., ZOU, Y., MARWAN, N., KURTHS, J. “Complex networks in climate dynamics”, *The European Physical Journal Special Topics*. 2009.
- HYDROREVIEW, Brasil, jan. de 2023. Disponível em: <<https://www.hydroreview.com/business-finance/hydropower-dominates-in-brazil-at-62-17-of-total-installed-capacity-at-end-of-2020/#gref>>.
- INTRODUCTION — NetworkX 1.7 documentation, outubro de 2023. Disponível em: <[Introduction — NetworkX 1.7 documentation](#)>.
- LELIS, T.A., CALIJURI, M.L., FONSECA, A.S., LIMA, D.C. “Impactos causados pelas mudanças climáticas nos processos erosivos de uma bacia hidrográfica: Simulação de cenários”, *Revista Ambiente & Água*, v.6, n.2, 2011.
- RODRIGUES, F.A. and Costa, L. F. “Generalized connectivity between any two nodes in a complex network”, *Physical Review E*. 81. 2010.

WebODM para o Processamento de Imagens Aéreas Obtidas por Drone em uma Área Rural: como alternativa aos softwares proprietários

Felipe Abraham Ramalho Xaud¹, Francisco Carlos de Lima Pereira², Clenya Rejane Barros de Lima², Wender Antônio da Silva², Haron Abraham Magalhães Xaud³

¹Curso de Ciência da Computação – Universidade Estadual de Roraima (UERR)
Boa Vista – RR - Brasil

²Colegiado de Ciência da Computação – Universidade Estadual de Roraima (UERR)
Boa Vista – RR – Brasil

³Empresa Brasileira de Pesquisa Agropecuária (Embrapa Roraima)
Boa Vista – RR – Brasil

felipexaude@gmail.com, fpcarlos@uerr.edu.br, clenyabarro@uerr.edu.br,
wender@uerr.edu.br, haron.xaud@embrapa.br

Abstract. *Drones and their various applications have been gaining great space around the world. In remote sensing, drones excel in collecting information with high resolution and accuracy. To generate these information, it is necessary to process the images with specific photogrammetry software. However, most of these software are proprietary with high prices in restrictive licenses, thus making it difficult for some groups to access this technology. This work aims to evaluate the free software WebODM to process images obtained by drone in a rural area, as an alternative to proprietary software. Our results have shown that WebODM can be used as a feasible alternative to obtain orthomosaics and DEM as high-quality products to photointerpretation.*

Resumo. *Os drones e suas diversas aplicações vêm ganhando grande espaço ao redor do mundo. No sensoriamento remoto, os drones se destacam na coleta de informações com alta resolução e precisão. Para gerar essas informações, é necessário processar as imagens com software específico de fotogrametria. No entanto, a maioria desses softwares são proprietários com preços elevados em licenças restritivas, dificultando o acesso de alguns grupos a essa tecnologia. Este trabalho tem como objetivo avaliar o software livre WebODM para processamento de imagens obtidas por drone em uma área rural, como alternativa aos softwares proprietários. Nossos resultados mostraram que o WebODM pode ser utilizado como uma alternativa viável para obter ortomosaicos e DEM como produtos de alta qualidade para fotointerpretação.*

1. Introdução

Os Veículos Aéreos Não-Tripulados (VANTs), também conhecidos como Aeronaves Remotamente Pilotadas (ARPs), são popularmente conhecidas como *drones*. A utilização de drones vem crescendo nos últimos anos, devido à facilidade de acesso à essa

tecnologia, sua popularização e sua aplicação nas mais diversas áreas, como agricultura, segurança e construção civil. Os avanços tecnológicos proporcionaram a acoplagem de diversos equipamentos a esses veículos não-tripulados, desde câmeras de alta definição a pulverizadores agrícolas [Prudkin e Breunig 2019], fato esse que vem diversificando sua utilização.

No sensoriamento remoto, os drones se destacam na coleta de informações de alta resolução e precisão. Entre os softwares mais populares de processamento de imagens obtidas por drones atualmente, podemos citar os softwares proprietários Agisoft Metashape, o Pix4dMapper e o DroneDeploy. Todos eles consolidados no mercado. O grande problema está nos altos valores cobrados para o uso destes softwares, que podem ser inviáveis para diversos usuários, como profissionais autônomos, pequenas empresas, grupos de pesquisas e extensionista [Santos e Koenigkan 2018].

Em alternativa aos softwares proprietários existem softwares de processamento de imagens gratuitos e de código abertos, como por exemplo o WebODM do projeto OpenDroneMap. Este trabalho apresenta um comparativo entre o WebODM e os softwares proprietários Agisoft Metashape, Pix4dMapper e DroneDeploy.

O WebODM é um software livre de processamento de imagem fotogramétrico voltado para imagens aéreas obtidas por drones, mantido pelo projeto OpenDroneMap, criado por Piero Toffanin e Dakota Benjamin, e está sob a licença GNU AGPL versão 3. O repositório oficial do WebODM está na plataforma GitHub, e conta com mais de 100 desenvolvedores atualmente. O WebODM é uma interface gráfica para o usuário, desenvolvido em Django (framework web escrito em python) e React.js (biblioteca javascript).

2. Materiais e Métodos

As imagens obtidas por drone são imagens hipotéticas de uma área rural qualquer e foram cedidas por terceiros para essa pesquisa, a amostra é de uma área de aproximadamente 50 hectares (0,50 km²) com um cultivo agrícola cercado por uma floresta, e com um rio na parte sul. O modelo de drone utilizado na coleta das imagens foi o *DJI Phantom 4 Advanced* (Figura 1) com uma câmera de 20MP com sensor CMOS 1", dispositivo GNSS (GPS/GLONASS). Utiliza o Sistema Geodésico de Referência (SGR) WGS 84. Um drone de entrada de captura de imagens, com asas rotativas, e tempo máximo de voo de 30 minutos.

2.1 Estação de Trabalho e Escolha dos Softwares

A estação de trabalho responsável pelo processamento é composta por processador i7-9750H com 6 núcleos (hexa-core), 12 threads com clock máximo de 4.50 GHz, memória RAM de 16GB (2x8GB) 2666MHz DDR4, armazenamento SSD SATA de 1TB e uma placa de vídeo Nvidia GeForce GTX 1660Ti 6GB GDDR6. Dotado dos sistemas operacionais Windows 10 e Ubuntu 22.04. Para selecionar quais softwares seriam usados na pesquisa foi feito um levantamento bibliográfico, documental e de mercado dos softwares mais utilizados nesse tipo de processamento de imagem. Definiu-se o uso dos softwares proprietários: Agisoft Metashape Professional (v. 1.8.3), o Pix4dMapper (v. 4.8.0) e o DroneDeploy (v. 2.162.0); todos em modo de teste gratuito ("full version"). O software livre utilizado foi o WebODM (v. 1.9.13).

2.2 Método Avaliativo

Para avaliação foram selecionadas 66 imagens, um total de 484 MB (megabytes), com uma sobreposição longitudinal de 70% e latitudinal de 50%, submetidas a um processamento de nível médio, que na maioria dos softwares é o nível padrão (default), tentando se chegar a um GSD (Ground Sample Distance) de 5,28 cm/pixel a 6,6 cm/pixel para o ortomosaico e um GSD de até 21.4cm/pixel para o DEM (Digital Elevation Model). A avaliação foi dividida em duas partes conforme fluxograma da Figura 1.

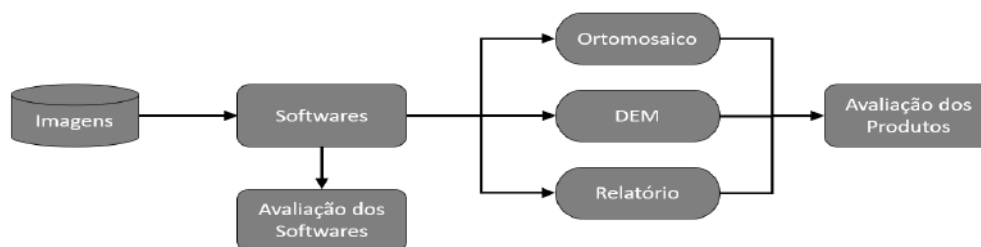


Figura 1. Fluxograma da metodologia

2.2.1 Método para avaliação dos softwares

Nessa avaliação foram explorados os aspectos técnicos de cada software com base na documentação e nos processamentos de imagem, tais como: 1. Sistema Operacionais Compatíveis; 2. Requisitos mínimos de memória RAM; 3. Requisitos mínimos de Espaço Livre em Disco; 4. Requisitos mínimos de CPU; 5. Faz uso da GPU; 6. Requisitos recomendáveis de memória RAM; 7. Processa em Nuvem; 8. Quantidade de extensões suportadas; 9. Preço; 10. Tempo de Processamento das 66 imagens; 11. Quantidade de imagens alinhadas. Além desses itens, também foi explorado um pouco da arquitetura e ecossistema dos softwares.

2.2.2 Método para avaliação dos produtos gerados

Os ortomosaicos e DEMs gerados por cada software foram submetidos à uma avaliação de um especialista fotointérprete voluntário da área florestal, com doutorado em sensoriamento remoto pelo INPE e experiência de mais de 25 anos em fotointerpretação. Para analisar os produtos foram utilizados elementos de fotointerpretação: tonalidade/cor, textura, tamanho, forma, sombra, altura, localização, padrão e contexto. Não se tratando da realização de um mapeamento em si, mas de uma avaliação prévia do potencial de uso dos produtos gerados para um posterior mapeamento dos principais alvos e objetos presentes na área amostral. As feições examinadas nas imagens foram: 1. Solo exposto; 2. Água/Rio/Lago; 3. Sombras; 4. Estradas; 5. Vegetação rasteira; 6. Arbustos/Arvoretas isolados; 7. Palmeiras; 8. Árvores isoladas; 9. Dossel florestal; 10. Árvores emergentes; 11. Coerência geral com o campo. Para análise visual dos produtos gerados foi utilizado o software livre QGIS (v. 3.10).

3. Resultados e Discussões

3.1 Comparações dos Softwares

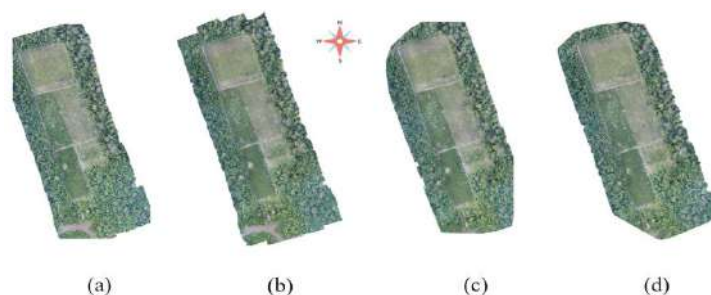
O WebODM e o Metashape apresentaram um maior tempo de trabalho até se conseguir os produtos, isso se deve ao processo custoso de instalação manual do WebODM, que pode ser contornado comprando o instalador para Windows. No caso do Metashape isso se deve ao seu fluxo de trabalho com maior nível de complexidade como aponta [Talib et al. 2021], esse tempo também diminui com a experiência do usuário na utilização do software. O DroneDeploy apresentou menor tempo de trabalho dos 4 softwares, por não precisar de instalação e ter uma usabilidade fácil. Seguido do Pix4dMapper, com um tempo de trabalho um pouco maior pois precisa ser instalado, mas também possui um fluxo de trabalho simples.

Em relação a compatibilidade com os Sistemas Operacionais, o WebODM e o Metashape se destacam por serem compatíveis com Linux, MacOS e Windows. O Pix4dMapper atualmente só pode ser instalado no Windows 10, fato que limita a quantidade de usuários. E, em se tratando da quantidade de extensões suportadas, o Metashape e o Pix4dMapper se destacam por suportarem mais de 20 extensões cada, quantidade superior aos outros dois softwares. Uma limitação do WebODM é que ele ainda não está otimizado para fazer uso de algumas placas de vídeo dedicadas (GPUs), houve a tentativa de configurar seu uso no decorrer dessa pesquisa, com a ajuda de fóruns da própria comunidade OpenDroneMap, mas nos testes o uso da GPU pelo WebODM não passou de 30%. No que se diz respeito a transparência do fluxo de trabalho, o WebODM se destaca por ser software livre [Von Flach e Kon 2021], o acesso ao código fonte e, conseqüentemente, sua arquitetura fornecem clareza e controle sobre os dados de todas as etapas do processamento e ferramentas utilizadas pelo software, como bibliotecas, linguagem de programação e banco de dados.

O tempo de processamento das 66 imagens coletadas foi de aproximadamente 15 minutos com o DroneDeploy, de 8 minutos com o Metashape, de 16 minutos com o Pix4dMapper e de 10 minutos com o WebODM.

3.2 Avaliação dos Produtos Gerados

Avaliando a qualidade dos ortomosaicos na Figura 2 nota-se que o de maior qualidade foram os produtos gerados pelo Metashape (b). Os ortomosaicos gerados pelo DroneDeploy (a), Pix4dMapper (c) e pelo WebODM (d) sofreram algumas perdas de bordas em diferentes níveis, mesmo assim, ainda satisfazem o uso para essa avaliação.



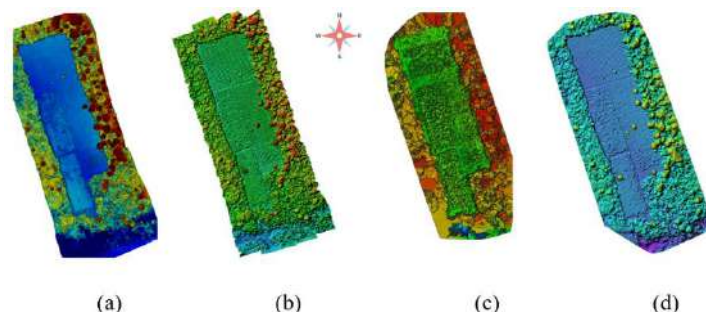


Figura 2. Ortomosaicos e DEMs: (a)DroneDeploy, (b)Metashape, (c)Pix4dMapper, (d)WebODM

Com ajuda do especialista em sensoriamento remoto, os ortomosaicos e DEMs foram avaliados e os resultados podem ser vistos na Tabela 1.

Com base nos dados interpretados da Tabela 2, é possível perceber que os softwares Metashape e WebODM apontaram vantagem aos demais softwares no item 8 (Árvores isoladas). Nos demais itens todos apresentaram uma boa separação, com exceção do item 10 (Árvores Emergentes), onde todos os softwares apresentaram uma separação mediana. De acordo com os dados apresentados, nota-se uma desvantagem do software DroneDeploy na separação referente ao item 8 (Árvores isoladas), onde esse apresentou uma separação de baixo nível, em comparação aos demais softwares. Outro detalhe refere-se ao item 6 (Arbustos/Arvoretas isolados), onde os softwares Metashape e Pix4dMapper conseguiram fazer uma separação mediana, e os demais softwares não separaram bem. No mais, observa-se um grande crescimento e amadurecimento do projeto OpenDroneMap. Burdziakowski (2017) em uma versão inicial do módulo ODM (v. 0.3), o considerou como uma ferramenta promissora para o desenvolvimento de pesquisas e aplicações científicas.

Tabela 1 – Avaliação dos Ortomosaicos por Fotointerpretação

Elemento de Fotointerpretação	Orto-DroneDeploy	Orto-Metashape	Orto-Pix4D	Orto-WebODM
1. Solo exposto	(+++)	(+++)	(+++)	(+++)
2. Água/Rio/Lago	(+++)	(+++)	(+++)	(+++)
3. Sombras	(+++)	(+++)	(+++)	(+++)
4. Estradas	(+++)	(+++)	(+++)	(+++)
5. Vegetação rasteira	(+++)	(+++)	(+++)	(+++)
6. Arbustos/Arvoretas isolados	(+++)	(+++)	(+++)	(+++)
7. Palmeiras	(+++)	(+++)	(+++)	(+++)
8. Árvores isoladas	(+ -)	(+++)	(+ -)	(+++)
9. Dossel florestal	(+++)	(+++)	(+++)	(+++)
10. Árvores Emergentes	(+ -)	(+ -)	(+ -)	(+ -)
11. Coerência geral com campo	(+++)	(+++)	(+++)	(+++)

Legendas: (+++) Separa Bem; (+ -) Separa Medianamente; (---) Não Separa Bem

Tabela 2 – Avaliação dos DEMs por Fotointerpretação

Elemento de Fotointerpretação	DEM - DroneDeploy	DEM - Metashape	DEM - Pix4D	DEM - WebODM
1. Solo exposto	(---)	(---)	(---)	(---)
2. Água	(---)	(---)	(---)	(---)
3. Sombras	(---)	(---)	(---)	(---)
4. Estradas	(---)	(---)	(---)	(---)
5. Vegetação rasteira (Gramíneo-lenhosa)	(---)	(---)	(---)	(---)
6. Arbustos/Arvoretas isolados	(---)	(+ -)	(+ -)	(---)
7. Palmeiras	(---)	(---)	(---)	(---)
8. Árvores isoladas	(---)	(+++)	(+++)	(+++)
9. Dossel florestal	(+++)	(+++)	(+++)	(+++)
10. Árvores Emergentes	(+++)	(+++)	(+++)	(+++)
11. Coerência geral com o campo	(+++)	(+++)	(+++)	(+++)

Legendas: (+++) Separa Bem; (+ -) Separa Medianamente; (---) Não Separa Bem

4. Considerações Finais

O Agisoft Metashape demonstrou melhor qualidade nas imagens do ortomosaico e do DEM, comparando com os demais softwares. No entanto, a automação no fluxo de trabalho do Pix4dMapper e do DroneDeploy foram notórias, justificando a popularidade desses softwares. O software livre WebODM atendeu às demandas comparativas aos softwares proprietários, em relação à alta qualidade do ortomosaico e do DEM, agregando todas as vantagens do uso de software livre. Contudo, há ainda algumas limitações como a incompatibilidade com algumas GPUs e o baixo número de extensões suportadas.

Referências

- Burdziakowski, P. (2017) “Evaluation of Open Drone Map toolkit for geodetic grade aerial drone mapping”, In: 17th International Multidisciplinary Scientific GeoConference SGEM, p. 101-110.
- Prudkin, G. e Breunig, F. (2019) Drone e Ciência: Teoria e aplicações metodológicas, vol. 1, FACOS-UFSM, RS, Santa Maria.
- Santos, T. e Koenigkan, L. (2018) Produção de Ortomapas com VANTs e OpenDroneMap, Circular Técnica, Embrapa, Campinas.
- Talib, N., Saad, M. S. M., Aminah S., Anshah, S. A., Saad N. M., Ghazali M. D., Abdullah M. A. e Jailani W. J. (2021) “A Comparative Flight Altitude Using Multi-Rotor UAV for Oil Palm Tree Counting”, In IOP Conference Series: Earth and Environmental Science, 2021.
- Von Flach, C. e Kon, F. (2021) “Software Livre: pré-requisito para ciência aberta”, In. Revista Computação Brasil, n. 46, SBC, Porto Alegre.

FMASK algorithm applied to Landsat images: a study case in the region of Santarém, Pará, Brazil

Elton Vicente Escobar-Silva^{1*}, Anielli Rosane de Souza^{1*}, Maria Lucia Ferreira Barbosa^{1*}, Cláudia Maria de Almeida¹, Liana Oighenstein Anderson², Gilberto Ribeiro de Queiroz¹ and Thales Sehn Körting¹

¹Earth Observation and Geoinformatics Division, National Institute for Space Research (INPE), São José dos Campos, São Paulo, Brazil

{elton.silva, anielli.souza, maria.lucia, claudia.almeida, gilberto.queiroz, thales.korting}@inpe.br

²National Center for Monitoring and Early Warning of Natural Disasters (Cemaden), São José dos Campos, São Paulo, Brazil

liana.anderson@cemaden.gov.br

*These authors contributed equally to the work.

Abstract: Clouds and cloud shadows are recurrent problems in optical remote sensing images. Algorithms have been developed for automatic detection of these targets such as the Function of Mask (Fmask) model. This work presents an automatic mapping of cloud areas with Fmask in Landsat images between 2017 and 2018 in northern Santarém, Pará, Brazil. The results show that the algorithm presents a good performance (85%) in the study area and is a powerful tool in clouds and cloud shadows detection. However, for further works, it is suggested to use data at a larger time scale and the evaluation of more pixels within the images, when comparing Fmask with other algorithms.

1. Introduction

Optical remote sensing images have a variety of applications that allow for new research areas in scientific research. However, considering land use and land cover (LULC) studies, even with some advances and upgrading, optical sensors are affected by clouds and cloud shadows, a problem in the visible and infrared spectral regions. These image effects limit the monitoring of LULC by obstructing targets, reducing the useful area of analysis, and interfere with the image quality itself (Jensen, 2009; Rees, 2013).

Automatic methods for mapping clouds and cloud shadows have been developed taking into account different approaches. Some studies rely on radiometric and color attributes to detect clouds and cloud shadows (Polidorio et al., 2006). Others are based on thresholds (Cucu-Dumitrescu, 2013) and on object analysis, such as the Function of mask (Fmask) method (Zhu and Woodcock, 2012), which is widely used both to eliminate unwanted noise in the pre-processing stage and to study the objects resulting from the data output (Pletsch et al., 2018; Foga et al., 2017; Braaten et al., 2015).

Fmask automatically extracts cloud, cloud shadows, snow, and water information from satellite images. The algorithm was developed by Zhu and Woodcock (2012) and its first version was developed only for Landsat-4 to 7 series images.

However, in 2015, Zhu et al. (2015) released a new version with some improvements and the algorithm was expanded to Landsat-8 and Sentinel-2 images. Fmask can be not only applied in the pre-processing of images, to minimize the inclusion of undesired targets, but also in the result analysis (Pletsch et al., 2018), which makes it a powerful tool for image analysis in areas with high recurrence of clouds.

The Amazon, located in the intertropical convergence zone (ITCZ), is characterized by a dense presence of clouds. Thus, the presence of clouds can be a major limiting factor in the use of remote sensing images in this region. Considering the great importance of the Amazon in the context of climate change and increasing pressures to change land cover, the development and validation of a tool aimed for cloud mask is essential for improvements in the use of optical images in this region, since the oversight of these areas are impacted by the clouds. In this context, the objective of the present study is (i) to apply the Fmask algorithm to detect clouds in Landsat images in the northern portion of Santarém, Pará; (ii) to analyze the number of times a pixel is classified as cloud; and (iii) to present the classification variation of the same pixel over time.

2. Materials and Methods

2.1 Study area

The area under analysis encompasses the 227/62 path/row (Figure 1) of Landsat 8. The region is located in the '*Baixo Amazonas*' mesoregion, northern Santarém in the Legal Amazon, and it includes part of the municipalities of Belterra, Mojuí dos Campos, Monte Alegre, Prainha, Santarém, and Uruará. Moreover, the area has different land regimes such as Protected Areas, INCRA (National Institute for Colonization and Agrarian Reform) settlement projects, indigenous lands, and agro-extractive settlement projects (IBGE, 2014; Souza et al., 2017).

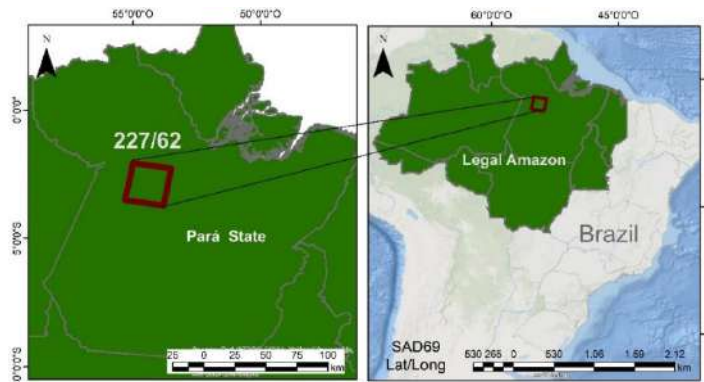


Figure 1: Geographic location of the study area.

2.2 Materials

Landsat has a totally free of charge database available on the United States Geological Survey (USGS) Earth Explorer website. On this website, images were acquired from the Operational Landsat Imager (OLI) sensor and from the Thermal Infrared Sensor (TIRS) Level-1 with radiometric and geometric corrections. The time series imagery extends over 2017 and 2018, totaling 24 images. Aiming to keep homogeneous weather

conditions throughout the time series, we selected images acquired around the same day in each month along the two analyzed years.

2.3 Methods

The methodology is structured into two major steps. The first step (Section 2.3.1) comprises the execution of Fmask and the second step (Section 2.3.2) describes the methodological approach for analyzing the algorithm results.

2.3.1 Fmask

In this study, the Python-Fmask version 0.5.4 package was used, executed in the Python language version 3.8.2. The Python code can be downloaded [here](#). To map the clouds and cloud shadows, digital number (DN) Landsat-8 images were converted into top of the atmosphere (TOA) reflectance images by the package itself. It is worth noting that the most recent version of the Fmask package accepts both TOA and DN images as input. If one chooses to use the DN images, it is important to verify if the file containing the images metadata ('MTL.txt') is located in the same directory of the images themselves (Flood and Gillingham, 2015). Table 1 describes the spectral bands of the employed images.

Table 1. Spectral bands of Landsat-8 images used in this work.

OLI (μm)	TIRS (μm)
Band 1 (0.43–0.45)	Band 10 (10.60–11.19)
Band 2 (0.45–0.51)	Band 11 (11.50–12.51)
Band 3 (0.53–0.59)	
Band 4 (0.64–0.67)	
Band 5 (0.85–0.88)	
Band 6 (1.57–1.65)	
Band 7 (2.11–2.29)	
Band 8 (0.50–0.68)	
Band 9 (1.36–1.38)	

The main purpose of Fmask is mapping clouds, cloud shadows, and snow. However, with the algorithm improvements, the resulting data is a raster file containing six classes: (1) null; (2) no-cloud; (3) cloud; (4) cloud shadow; (5) snow; and (6) water. In this study, null values were removed, and the images were reclassified again from 1 to 5, starting with the class no-cloud and ending with water.

Lastly, this algorithm identifies pixels with a high probability of being clouds according to the physical properties of a cloud, separating them from other targets on the Earth's surface. The Python-Fmask package used in this work is structured in a high-level Python programming language. A concise description of the package can be found in Flood and Gillingham (2015). In the command terminal, the Fmask package was executed for all 24 images in the analyzed period.

2.3.2 Classification analysis

Initially, the frequency at which a pixel was classified as cloud in the study period was assessed. This analysis was divided into two stages. In the first one, a qualitative

analysis was carried out through visual interpretation, where images from the Landsat series were compared to the classification result, in order to observe the correspondence between areas mapped as cloud and no-cloud areas in the image. A vertical bar graph was also created using the matplotlib library, considering the total pixels mapped in each class for July 2018 as an experimental period. Furthermore, 100 random points were generated in QGIS and visually checked if they were correctly classified. In the second stage, a temporal analysis at a pixel level was performed, where it was possible to observe the class changes within the analyzed period. The raster values were extracted over time from a random pixel centroid corresponding to the Fmask classes. For visualization purposes, bar graphs were plotted using the matplotlib.

3. Results

Fmask presented a satisfactory performance in the detection of clouds in the study area. By visual comparison, it was observed that the algorithm correctly classified the resulting classes (no-cloud, cloud shadow, snow, and water). 85% of randomly generated points were classified correctly. Most of the error is associated with the classification of the pixel with cloud or shadow when in fact the pixel is neither related to these two classes. According to Qiu et al. (2019), there is no optimal global threshold for cloud detection, so local thresholds should be applied by users according to a given region's peculiarities. Furthermore, accurately detecting cloud shadows is a major challenge as many dark surfaces (e.g., wetlands, burnt areas and terrain shadows) are easily mistaken for cloud shadows (Zhu et al. 2015; Qiu et al. 2019).

Considering the Fmask result for July 2018, the dry season, it was observed that the classification generated accurate results when compared to the Landsat image (Figure 2). In July 2018, the water class represented 11.80% of the total pixels of the classified image (Figure 3). This class is associated with the Tapajós, Arapiús, and Amazon rivers, which were correctly identified by the algorithm. The cloud shadow class was registered as 7.80% of the total pixels. The cloud and cloudless classes, the largest classes in the image, represented 14.24% and 66.14% of the pixels, respectively. Finally, no pixels belonging to Class 4 (snow) were identified, indicating a good accuracy of the algorithm since this class does not occur in the study area.

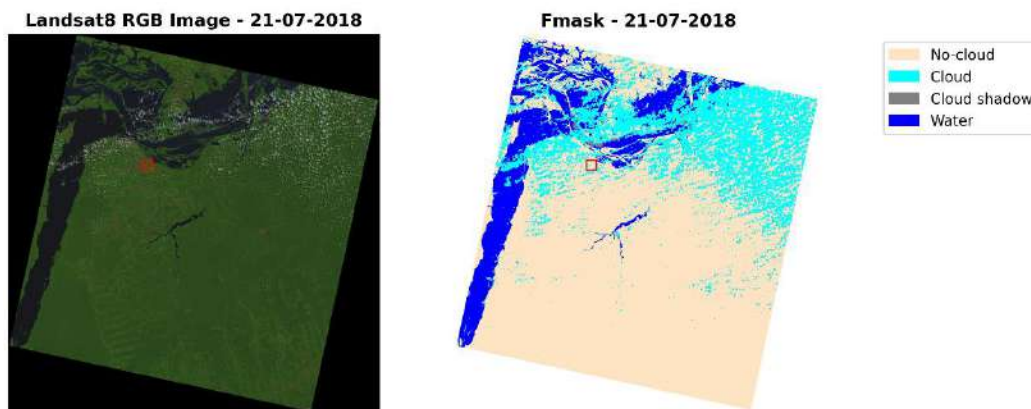


Figure 2: RGB Landsat 8 image (left) and Fmask (right) for the same scene on 17 Jul 2018. The red box on both images represents the centroid location of the analyzed pixel in this work.

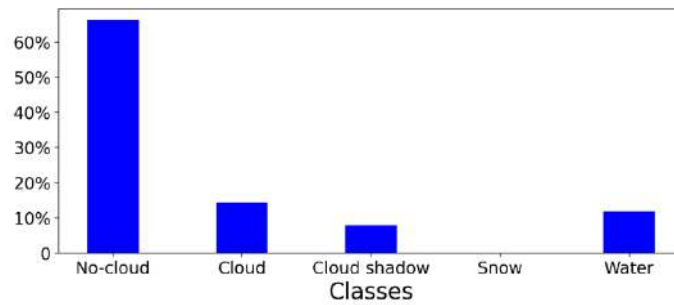


Figure 3: Percentage of mapped pixels per class in July 2018.

The class change corresponding to the analysis of a pixel over the 24 images (2017 and 2018) is presented in Figure 4. In this case, the pixel was identified within only two classes, with cloud and no-cloud, Classes 1 and 2, respectively. It is possible to identify that the pixel was identified as cloud in most months, even during the dry season in the two analyzed years. This pattern could be explained by the low representation of a pixel when compared to the full scene.

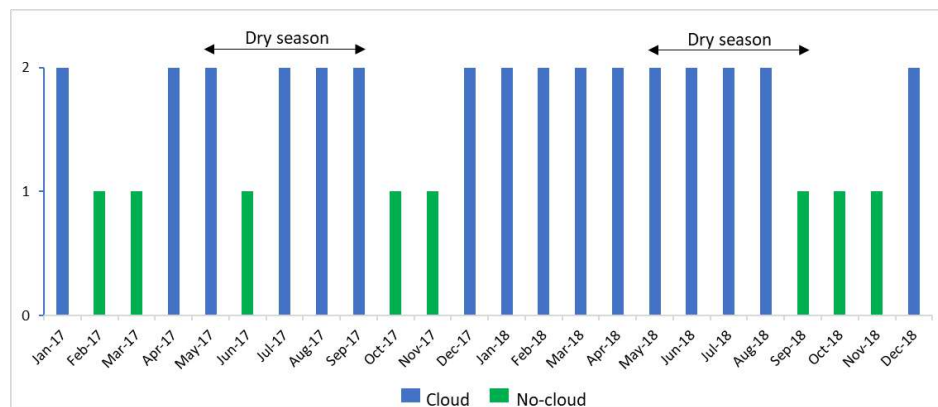


Figure 4: The class variation between 2017 and 2018 of the selected pixels. See pixel location in Figure 2.

In 2017, the analyzed pixel was identified as cloud in 58.3% of the months and as no-cloud area in 41.7%. In 2018, the cloud proportion raised to 75%, while the share of no-cloud classification dropped to 25%. When observing the two years altogether, in 66.7% of the time series, the pixel was registered as cloud, and in 33.3% of it as no-cloud. Lastly, as expected, the algorithm identified a greater presence of clouds in the images referring to the rainy season (October to April) and a greater presence of cloudless areas in the dry season (May to September).

4. Final considerations

The Fmask algorithm showed a good performance in the study area (85% of accuracy). The snow class, which does not occur in the Amazon, was not computed by the algorithm, indicating a reliable result. However, the algorithm reduces the image size by removing some columns in the image edges, which can cause a small loss of image information. As directions for future work, it is recommended to compare the performance of Fmask with other cloud detection algorithms in the same study area.

5. Acknowledgments

E.V.E.S. thanks the São Paulo Research Foundation (FAPESP, Grant 2021/11435-4). E.V.E.S., A.R.S. and M.L.F.B. thank the National Council for Scientific and Technological Development (CNPq, Funding Code 001). C.M.A. thanks FAPESP (Grant 2020/09215-3) and CNPq (Grant 311324/2021-5).

6. References

- BRAATEN, Justin D.; COHEN, Warren B.; YANG, Zhiqiang. 2015. Automated cloud and cloud shadow identification in Landsat MSS imagery for temperate ecosystems. *Remote Sensing of Environment*, 169, pp.128-138.
- CUCU-DUMITRESCU, Catalin. 2013. A simple method of determining cloud-masks and cloud-shadow-masks from satellite imagery. *IEEE Geoscience and Remote Sensing Letters*, 11(1), pp.10-13.
- FLOOD, N.; GILLINGHAM, S. 2015. *Python Fmask documentation*. Disponível em <<http://www.pythonfmask.org/en/latest/index.html>> Acessado em Jun 06 2022.
- FOGA, Steve; SCARAMUZZA, Pat L.; GUO, Song; ZHU, Zhe; DILLEY JR, Ronald D.; BECKMANN, Tim; SCHMIDT, Gail L.; DWYER, John L.; HUGHES, M. Joseph; LAUE, Brady. 2017. Cloud detection algorithm comparison and validation for operational Landsat data products. *Remote sensing of environment*, 194, pp.379-390.
- IBGE - Instituto Brasileiro de Geografia e Estatística. *Redes e Fluxos do Território: Gestão do Território*. 2014.
- JENSEN, John R., 2009. *Remote sensing of the environment: An earth resource perspective*. 2nd ed. Pearson Education India.
- PLETSCH, Mikkaela A. J. S.; PENHA, T. V.; SILVA JUNIOR, C. H. L.; KÖRTING, Thales S.; ARAGÃO, Luiz E. O. e C.; ANDERSON, Liana O. 2018. Integração do algoritmo fmask ao modelo linear de mistura espectral como subsídio à detecção de áreas queimadas na Amazônia brasileira. *Revista Brasileira de Cartografia*, vol. 70, pp. 696–724.
- POLIDORIO, Airton M.; IMAI, Nilton N.; TOMMASELLI, Antonio M. G.; FLORES, Franklin C. 2006. Detection and Discrimination of Shades, Clouds and Waterbodies in Remote Sensing Images. *Revista Brasileira de Cartografia*, vol. 58, pp. 223-232.
- QIU, Shi; ZHU, Zhe; HE, Binbin. (2019). Fmask 4.0: Improved cloud and cloud shadow detection in Landsats 4–8 and Sentinel-2 imagery. *Remote Sensing of Environment*, 231, p.111205.
- REES, William G. 2013. *Physical Principles of Remote Sensing*. 4th ed. Cambridge: Cambridge University Press.
- SOUZA, Anielli R.; ESCADA, Maria I. S.; VIEIRA, Antônio M. V. 2017. Padrão da paisagem associado ao uso e cobertura da terra em comunidades ribeirinhas e de terra firme situadas no sudoeste do Pará. *Geografia*, vol. 42, no. 2, pp. 135–164.
- ZHU, Zhe; WANG, Shixiong; WOODCOCK, Curtis E. 2015. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote sensing of Environment*, 159, pp.269-277.
- ZHU, Zhe; WOODCOCK, Curtis E. 2012. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote sensing of environment*, 118, pp.83-94.

,

Controle da Qualidade Cartográfica de *Basemaps*: Avaliação da Completude

Marconi Martins Cunha¹, Lígia da Silva Barbosa¹, Isaac Antunes Miranda Junior¹, Afonso de Paula dos Santos¹

¹Departamento de Engenharia Civil – Universidade Federal de Viçosa (UFV)
Av. Peter Henry Rolfs, s/n - Campus Universitário – 36570-900 – Viçosa – MG – Brasil

{marconi.cunha, ligia.barbosa, isaac.junior, afonso.santos}@ufv.br

Abstract. *This work aimed to evaluate the completeness of the OpenStreetMap, Google Road, Bing Map and CARTO Voyager basemaps in the central area of the Federal University of Viçosa (Viçosa-MG), considering nine layers, and the omission and commission subelements presented by ET-CQDG. With the results obtained, a non-conforming result was found in relation to the omission in all layers in all basemaps, which did not occur in relation to the commission, since in some layers some basemaps presented a conforming result. Thus, it was noted the need to improve the information mainly absent from these basemaps in these areas, aiming at a better representation of the real world.*

Resumo. *Este trabalho teve como objetivo avaliar a completude dos basemaps OpenStreetMap, Google Road, Bing Map e CARTO Voyager na área central da Universidade Federal de Viçosa (Viçosa-MG), considerando nove camadas, e os subelementos omissão e comissão apresentados pela ET-CQDG. Com os resultados obtidos, constatou-se um resultado não conforme em relação à omissão em todas as camadas em todos os basemaps, o que não ocorreu em relação à comissão, visto que em algumas camadas alguns basemaps apresentaram um resultado conforme. Dessa forma, notou-se a necessidade de melhoria das informações principalmente ausentes desses basemaps nessa área, visando uma melhor representação do mundo real.*

1. Introdução

Nas últimas décadas, com o desenvolvimento da *internet*, aumentou-se a disponibilidade de informações espaciais na *internet*. Um exemplo desta disponibilidade são os *basemaps*. Os *basemaps* são compostos por imagens originadas de sensores em satélites ou aerotransportados, e podem servir de base ou como auxílio para diversos fins, como mapeamento de usos dos solos, planejamento e gestão urbana, mapas temáticos, entre outros (SZTUTMAN, 2014). Entretanto, apesar de plataformas como estas serem amplamente utilizadas, a qualidade dos seus dados disponibilizados é desconhecida e/ou não informada aos usuários (ARIZA-LÓPEZ et al., 2020).

Nesse sentido, o Controle de Qualidade Cartográfica (CQC) se apresenta como um relevante processo para a investigação da qualidade de dados geoespaciais. Sendo que a avaliação desta é realizada por meio dos elementos de qualidade dos dados espaciais. De acordo com a ISO (2013), os elementos de qualidade são: completude, consistência lógica, acurácia posicional, acurácia temática, acurácia temporal e usabilidade.

A completude pode ser definida como a presença e ausência de feições, seus atributos e relacionamentos (ISO, 2013). Por lidar com a presença e/ou ausência de feições, a completude pode ser afetada por regras de generalização cartográfica adotada e a escala

dos dados (SANTOS, 2013). O processo de generalização cartográfica é executado visando estabelecer um equilíbrio entre a quantidade de dados representados e o foco do mapeamento, em que a escala tem significativa importância na quantidade de generalização a ser utilizada (ROBINSON et al., 1995; VIANNA, 1997). Portanto, é fácil perceber que essa alteração na quantidade de feições representadas dada pela generalização impacta diretamente a análise de completude (ARIZA-LÓPEZ, 2002).

Vale ressaltar que para avaliar os elementos de qualidade é fundamental a adoção de normas e padrões. No Brasil, as normas adotadas o Decreto nº 89.817 de 1984 (BRASIL, 1984) e a ET-CQDG (Especificação Técnica para Controle de Qualidade de Dados Geoespaciais) (DSG, 2016) da Infraestrutura Nacional de Dados espaciais (INDE).

Diante do exposto, o objetivo deste estudo consiste na avaliação da completude dos *basemaps* OpenStreetMap, Google Road, Bing Map e CARTO Voyager; através dos parâmetros presentes no Decreto nº 89.817 e na ET-CQDG.

2. Materiais e Métodos

2.1. Área de Estudo

A área de estudo deste trabalho (Figura 1) compreende a área central do *campus* Viçosa da Universidade Federal de Viçosa (UFV), localizada no município de Viçosa, Minas Gerais. Com uma extensão de aproximadamente 267 hectares, esta área é delimitada pelas latitudes 20°45'06" S e 20°46'24" S, e longitudes 42°51'06" W a 42°52'43" W.



Figura 1. Área de Estudo

Esta é uma área relativamente pequena, mas deve-se destacar que o objetivo principal deste trabalho é indicar uma metodologia, que possa ser reproduzida em áreas maiores.

2.2. Materiais

Os produtos avaliados neste estudo quanto a sua completude foram quatro *basemaps*: OpenStreetMap, Google Road, Bing Map e CARTO Voyager. Estes estão disponíveis no QGIS a partir do *plugin* INDE (PINHEIRO et al., 2020). Já as feições tidas como referência são provenientes da base cartográfica do *campus* Viçosa da UFV elaborada por Matta e Cunha (2017). Devido a essa disponibilidade dos dados, esta foi a área escolhida para ser avaliada. Além disso, foram utilizadas as imagens dos *basemaps* Esri Satellite e Bing Aerial para tirar possíveis dúvidas sobre a base de referência.

2.3. Métodos

Antes da análise da completude, foi realizada a análise da base cartográfica de referência. O espaço geográfico é dinâmico, sendo assim, uma base de dados criada em 2017 poderia

ter sofrido com desatualizações. Este fato é normal e, de fato, esperado. Entretanto, deve ser levado em consideração para que não interfira na avaliação da completude.

Nesta etapa foi realizado o levantamento de todas as feições que não estão mais presentes no mundo real ou que estão presentes, mas não se encontram representadas na base de referência. Isso foi feito com auxílio de imagens dos *basemaps* Esri Satellite e Bing Aerial e pelo conhecimento prévio da área de estudo por partes dos autores.

Um exemplo pode ser visto na Figura 2. Nesta imagem, onde as edificações estão representadas por polígonos pretos, na data da elaboração da base cartográfica, um prédio (em destaque azul em um *basemap* avaliado - Figura 2a - e verde na imagem do Bing Aerial - Figura 2b) havia sido demolido, por isso não se encontrava presente na base de dados. Entretanto, recentemente, foi construído outro prédio no lugar, motivo pelo qual aparece nos *basemaps*.

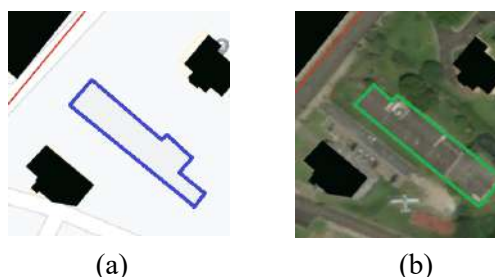


Figura 2. Exemplo de edificação faltante na base cartográfica de referência

Em situações como essa não foi contabilizado um erro de completude, por ser uma questão da referência, e não do produto avaliado. Após estas considerações, foram selecionadas 9 camadas (Tabela 1) cujos componentes seriam avaliados nos *basemaps*.

Tabela 1. Camadas utilizadas na avaliação da completude

Camada	Nome na base de referência	Nº de objetos de referência
Edifícios	ge_edf_edificacao_a,	259
	ge_edf_edif_religiosa_a,	
	ge_edf_edif_saude_a,	
	ge_edf_edif_constr_est_med_fen_a	
Campo/Quadra	ge_laz_campo_quadra_a	11
Piscina	ge_laz_piscina_a	2
Pista de Atletismo	ge_laz_pista_competicao_a	1
Pista de Pouso	pe_aer_pista_ponto_pouso_a	1
Massa d'água	pe_hid_trecho_massa_dagua_a	5

Para a avaliação da completude foram utilizadas as medidas porcentagem de itens em excesso (comissão) e porcentagem de itens ausentes (omissão) da ET-CQDG (DSG, 2016). Cabe destacar que a ET-CQDG apresenta outras medidas para avaliação da completude, mas somente a comissão e a omissão são aplicáveis aos *basemaps*.

De acordo com a ET-CQDG, a comissão é a razão entre o número de objetos que foram incluídos no conjunto avaliado, mas que não existem no conjunto de referência, e o número de objetos que deveriam realmente estar presentes (DSG, 2016). A comissão pode ser calculada por meio da Equação (1).

$$Comissão = \frac{\text{itens em excesso}}{\text{tamanho da referência}} \quad (1)$$

DSG (2016) também define a omissão, que é dada pela razão entre o número de objetos ausentes no conjunto avaliado e o número de objetos que deveriam realmente estar presentes. A omissão pode ser obtida por meio da Equação (2).

$$Omissão = \frac{\text{itens em falta}}{\text{tamanho da referência}} \quad (2)$$

Nos procedimentos para a avaliação dos conjuntos de dados geoespaciais vetoriais (CDGV) da ET-CQDG, este documento aponta que a taxa de erro para a completude, tanto para omissão quanto para comissão, deve ser menor que 4% para que o produto avaliado seja conforme. Sendo assim, se a comissão ou a omissão foi superior a 4%, o *basemap* foi considerado não conforme para aquelas feições avaliadas.

3. Resultados

A avaliação da completude dos quatro *basemaps* apontou 306 inconsistências em relação à base cartográfica de referência. Dentre estes, a grande maioria foram erros de omissão (301). Sendo que os *basemaps* OpenStreetMap, Bing Map e CARTO Voyager foram os que apresentaram o maior número de erros: 74, 100 e 91, respectivamente. Os detalhes destes resultados podem ser observados na Tabela 2.

De fato, a comissão não foi um problema: somente o Google Road apresentou erros de comissão, e ainda assim com um valor menor que o limite de 4% (1,93% para as edificações). Portanto, todos os *basemaps* avaliados quanto a sua completude foram considerados conformes para o subelemento comissão.

Já com relação à omissão, considerando os edifícios, o *basemap* Google Road apresentou 4 vezes menos erros que os demais *basemaps*. Entretanto, apesar dessa menor taxa de erro, nenhum *basemap* apresentou conformidade com relação à completude desta classe de objetos.

Além disso, para todas as outras camadas, a omissão do Google Road foi de 100%. Dos *basemaps* da Google disponíveis no *plugin* da INDE, o Road é o que apresenta maior número de feições representadas. Entretanto, como pode ser visto na avaliação da completude, este deixou de representar elementos importantes como as massas d'água, que na área de estudo retratam as 5 represas do *campus* Viçosa, e a pista de pouso. Este resultado foi semelhante ao do *basemap* Bing Map, que apresentou omissão de 100% em 5 camadas, sendo conforme quanto à omissão para a representação da pista de pouso.

É importante destacar que cada *basemap* tem as suas características, focando em representar determinadas feições em detrimento a outras. Os *basemaps* que apresentaram 100% de omissão, provavelmente decidiram não representar tal classe de feição. Por outro lado, também deve-se destacar que, independentemente do foco do *basemap*, seria interessante estas plataformas digitais apresentarem estas informações como a presença de corpos d'água (massa d'água) e pistas de pouso.

Ainda com relação à avaliação da omissão, os melhores desempenhos foram apresentados pelos *basemaps* OpenStreetMap e CARTO Voyager, que apresentam conformidade em quatro e três camadas, respectivamente.

Uma questão observada foi com relação à distribuição espacial dos erros. Foi observado que, em todos os *basemaps* analisados, a maior ocorrência dos erros de completude aconteceu em regiões mais afastadas da via principal do *campus* Viçosa da UFV, na área

central da região de estudo. Este fato pode indicar que as plataformas digitais priorizam as áreas centrais das regiões representadas.

Tabela 2. Avaliação da completude dos *basemaps*

	Camada	Itens Omissão	Itens Comissão	Omissão (%)	Comissão (%)	Resultado Omissão	Resultado Comissão
OpenStreetMap	Edifícios	70	0	27,03	0,00	Não Conforme	Conforme
	Campo/Quadra	4	0	36,36	0,00	Não Conforme	Conforme
	Piscina	0	0	0,00	0,00	Conforme	Conforme
	Pista de Atletismo	0	0	0,00	0,00	Conforme	Conforme
	Pista de Pousso	0	0	0,00	0,00	Conforme	Conforme
	Massa d'água	0	0	0,00	0,00	Conforme	Conforme
Google Road	Edifícios	16	5	6,18	1,93	Não Conforme	Conforme
	Campo/Quadra	11	0	100,00	0,00	Não Conforme	Conforme
	Piscina	2	0	100,00	0,00	Não Conforme	Conforme
	Pista de Atletismo	1	0	100,00	0,00	Não Conforme	Conforme
	Pista de Pousso	1	0	100,00	0,00	Não Conforme	Conforme
	Massa d'água	5	0	100,00	0,00	Não Conforme	Conforme
Bing Map	Edifícios	81	0	31,27	0,00	Não Conforme	Conforme
	Campo/Quadra	11	0	100,00	0,00	Não Conforme	Conforme
	Piscina	2	0	100,00	0,00	Não Conforme	Conforme
	Pista de Atletismo	1	0	100,00	0,00	Não Conforme	Conforme
	Pista de Pousso	0	0	0,00	0,00	Conforme	Conforme
	Massa d'água	5	0	100,00	0,00	Não Conforme	Conforme
CARTO Voyager	Edifícios	79	0	30,50	0,00	Não Conforme	Conforme
	Campo/Quadra	11	0	100,00	0,00	Não Conforme	Conforme
	Piscina	1	0	50,00	0,00	Não Conforme	Conforme
	Pista de Atletismo	0	0	0,00	0,00	Conforme	Conforme
	Pista de Pousso	0	0	0,00	0,00	Conforme	Conforme
	Massa d'água	0	0	0,00	0,00	Conforme	Conforme

Diante destes resultados, pode-se concluir que todos os *basemaps* avaliados podem melhorar a representação de feições, tendo em vista que mesmo feições que seriam básicas apresentaram não conformidade. Além disso, provavelmente o nível de generalização cartográfica dos *basemaps* pode ter afetado os resultados obtidos, levando em consideração as suas características em relação à escolha dos objetos que devem estar presentes. Isso porque, geralmente, os *basemaps* possuem cobertura global, exigindo alto grau de generalização cartográfica da base de dados, o que pode impactar a presença ou ausência de determinadas feições, o que afeta a análise de completude.

4. Conclusão

Com este trabalho foi possível analisar a completude dos *basemaps* OpenStreetMap, Google Road, Bing Map e CARTO Voyager. Analisou-se nove camadas por meio dos subelementos comissão e omissão. Verificou-se que houve mais problemas em relação à ausência de informações (omissão) do que o excesso de informações (comissão) na área avaliada, visto que apenas um *basemap* apresentou um erro de comissão em relação à uma das camadas, mas com valor menor que a tolerância adotada de 4%.

Foi possível notar então que todos *basemaps* testados foram conformes em relação à comissão. Entretanto, em relação à omissão foram poucas camadas que obtiveram o resultado conforme, demonstrando a necessidade de atualização e avaliação ao longo do

tempo dos *basemaps* avaliados. Além disso, o *basemap* Google Road foi o único que não foi conforme em todas as camadas ao avaliar a omissão.

Deve-se ressaltar que foi avaliada uma área relativamente pequena porque era onde se tinha dados cartográficos de referência, indispensáveis para este tipo de avaliação. Apesar do tamanho da área, os resultados deste trabalho já fornecem um indicativo da qualidade dos *basemaps*. Além disso, a importância deste trabalho também reside em indicar a metodologia utilizada, para que esta possa ser reproduzida em áreas maiores, como cidades, estados ou países.

Dessa forma, recomenda-se para trabalhos futuros a avaliação do elemento completude de outros *basemaps* presentes no plugin da INDE, considerando áreas ainda maiores, visto que a avaliação foi limitada à área estudada.

Referências

- Ariza-López, F. J. (2002) *Calidad en la producción cartográfica*. Madrid: Ra-Ma.
- Ariza-López, F. J., González, P. B., Pau, J. M., Torres, A. Z., Pascual, A. F. R., Vergara, G. M., Balboa, J. L. G. (2020) “Geospatial data quality (ISO 19157-1): evolve or perish”, *Revista Cartográfica*, (100), p. 129-154.
- Brasil. (1984) “Decreto N° 89.817 de 20 de junho de 1984. Estabelece as Instruções Reguladoras das Normas Técnicas da Cartografia Nacional”. Brasil.
- DSG - Diretoria do Serviço Geográfico. (2016) “Especificação Técnica Para Controle de Qualidade de Dados Geoespaciais (ET-CQDG)”. Brasília.
- ISO. (2013) “ISO 19157: Geographic information - Data quality”. International Organization for Standardization.
- Matta, J. P. A., Cunha, M. M. (2017) “Geração e disponibilização na Web de uma base cartográfica planimétrica da Universidade Federal de Viçosa, campus Viçosa”. Trabalho de Conclusão de Curso. UFV, Departamento de Engenharia Civil, Viçosa.
- Pinheiro, Julierme Gonçalves; Vieira, Alen Costa; Almeida, Warley Rodrigues; Bueno; Luis Fernando. (2020). INDE: Plugin QGIS para Interoperabilidade de Dados. In: *II Simpósio Brasileiro de Infraestrutura de Dados Espaciais (SBIDE)*, Rio de Janeiro: IBGE, p. 129-130.
- Robinson, A. H.; Morrison, J. L.; Muehrcke, P. C.; Kimerling, A. J.; Guptil, S. C. (1995). “Elements of Cartography”. New York: John Wiley & Sons. 6. ed.
- Santos, A. S. (2013). *Geoestatística aplicada na avaliação da qualidade de dados geoespaciais*. M.Sc. Viçosa, MG. UFV.
- Sztutman, Paulo. (2014) “Análise da qualidade posicional das bases do Google Maps, Bing Maps e da ESRI para referência espacial em projetos em SIG: aplicação para o município de São Paulo”. 2014. Dissertação - Escola Politécnica, USP, São Paulo.
- Vianna, C. R. F. (1997). “Generalização Cartográfica em Ambiente Digital – Escala 1:250.000 a partir de Dados Cartográficos Digitais na Escala 1:50.000”. Dissertação de Mestrado, Instituto Militar de Engenharia, Rio de Janeiro.

Analysis of mobility network metrics in the spread of COVID-19 in the Metropolitan Region of Vale do Paraíba and Litoral Norte (RMVPLN)

Leticia da S. Cabral¹, Leonardo Bacelar L. Santos², Antonio Miguel Vieira Monteiro³

¹Institute of Science and Technology - Paulista State University (UNESP)
São José dos Campos – SP – Brazil

²National Center for Monitoring and Natural Disaster Alerts (CEMADEN)
São José dos Campos – SP – Brazil

leticia.cabral1516@gmail.com, santosl1516@gmail.com, miguel.monteiro@inpe.br

³National Institute of Spatial Researches (INPE)
São José dos Campos – SP – Brazil

Abstract. *In 2019, a pandemic of the so-called new coronavirus (SARS-COV-II) began, which causes the disease COVID-19. In a short time after the first case appeared, hundreds of countries began to register new cases every day. Mapping and analyzing the flow of people, regardless of the mode of transport, can help us to understand and prevent several phenomena that can affect our society in different ways. Graphs are complex networks made up of points and edges. The (geo)graphs are graphs with known spatial location and, in the case of our study, the edges represent the flow between them. The (geo)graphs proved to be a promising tool for such analyses. In the study region, municipalities that first registered their COVID-19 cases are also municipalities that have the highest mobility indices analyzed: degree, betweenness and weight of edges.*

1. Introduction

The World Health Organization declared the SARS-COV-II virus a pandemic on March 11, 2020. Almost two months later, on April 30, 2020, the so-called new coronavirus, which causes the disease COVID-19, had already been identified in 180 countries. On February 26, 2020, the Ministry of Health confirmed the first case of the disease in Brazil, and currently, in August 2021, there are more than 200 million cases and 4 million deaths worldwide from COVID-19 [Sanarmed 2021].

The rapid spread of the virus has been accentuated by the ease with which people cross borders in the globalized world in which we live. Countries, states and cities around the world are strongly connected through different modes of transport, facilitating the spread of contagious diseases in an accelerated and uncontrolled way. Therefore, it is necessary to understand the mobility phenomena and adopt control measures to contain the flow of people, causing the transmission network to be interrupted [FREITAS 2020a].

The Metropolitan Region of Vale do Paraíba and Litoral Norte (RMVPLN) is made up of 39 municipalities divided into five sub-regions. Containing a population of approximately 2.5 million people [EMPLASA 2016], the region has a great connection with state and federal highways, airports and universities, factors that influence the displacement of people to the region and that end up influencing the dissemination of COVID-19

[PERES]. In addition, there is the pendulum movement that occurs from home to work and can occur intercity, being of great importance in the study of the displacement of people [JARDIM].

There is strong social interaction between the cities of the RMVPLN with the urban network of the state of São Paulo, with the capital and other cities such as Campinas, which is currently considered one of the 15 metropolises in Brazil [IBGE 2020].

Recent works in the literature demonstrate that there are patterns between mobility and the spread of COVID-19. It is possible to notice that the disease diffusion model in the state of São Paulo occurred from the largest to the smallest cities, characterizing a cascade effect in which, hierarchically, the propagation walked from the large metropolises to the regional capitals and other cities in the countryside of São Paulo [FORTALEZA 2020]. In [FREITAS 2020b], topological indices are used to represent their relationship with the spread of the disease.

Mathematical graphs have several applications [BOAVENTURA Netto 2012]. Complex networks are a type of graph with topological properties. The use of complex networks in the study of mobility provides analysis of important indices for the understanding of local dynamics. The (geo)graphs [SANTOS 2017] are defined as graphs composed of nodes of known geographic location and edges that connect pairs of nodes and represent their spatial dependence.

2. Methodology

For the elaboration of the (geo)graphs, two codes were executed in the Google Collaboratory platform in the Python language, previously written and provided to this study. The input data used were the coordinates of the centroids of the study region, a flow matrix and a threshold, which defines which edges are strong enough to appear in the (geo)graph.

In addition to the longitude and latitude of the extracted centroids, a flow matrix was used as input for the first code. It was prepared from an Origin-Destination matrix generated with 2014 year flow data from Sao Paulo Metropolitan Transport Department. The matrix contains flow data by the individual mode and by the work reason of the RMVPLN connected and also to the municipalities of Belo Horizonte/MG, Campinas/SP, Rio de Janeiro/RJ and São Paulo/SP. As input data for the second code, we used the same flow matrix and a file with the threshold, which for this study was defined as 1 (one), what means that edges with a flow bigger than or equal to one are strong enough to appear in the (geo)graph.

During the pre-processing of the matrix, we removed the rows and columns with data from 'other municipalities', as these did not have a defined spatial location, being not consistent with the definitions of a (geo)graph. In the execution of the code, the matrix was symmetrized by the sum of the values above and below the main diagonal, in order to obtain the total flow of the connection pair, either the origin or destination municipality of the displacement.

As output data of the two codes, the metrics of the nodes and the shapefile with the nodes and edges of the (geo)graph were obtained and used for the elaboration of three maps of the RMVPLN. In each map, a different index was used in the symbology, namely degree centrality, betweenness and edge weight. The first two are given by the node

metrics and the last one is provided by the edge attribute table. In all three indices, a color scale was applied in the symbology, representing the highest indices in red and the lowest in green gradually.

In the constructed (geo)graph, there are nodes, which represent the centroid of the municipalities in the study area, connected by edges, which represent the inter-municipal flow between them. Degree centrality represents the number of connections of each node of the (geo)graph, that is, how many different municipalities a given municipality connects to. The intermediation (betweenness) indicates how central a node is within the network, that is, how much a municipality acts as an intermediary in the shortest path between two other municipalities. And, finally, the weight of the edges (weight) represents the intensity of the flow between the municipalities.

In addition, we prepared maps with the representation of the occurrence of the first case in each municipality according to the epidemiological week. For this, we used data downloaded from the Wesley Cota website containing the cases of each municipality. Only the first case of each one was filtered and using the epidemiological week of the occurrence as an attribute in the symbology, it was possible to apply a color scale and represent the municipalities that had the first case registered first in red and those that had the first case registered finally in light green gradually.

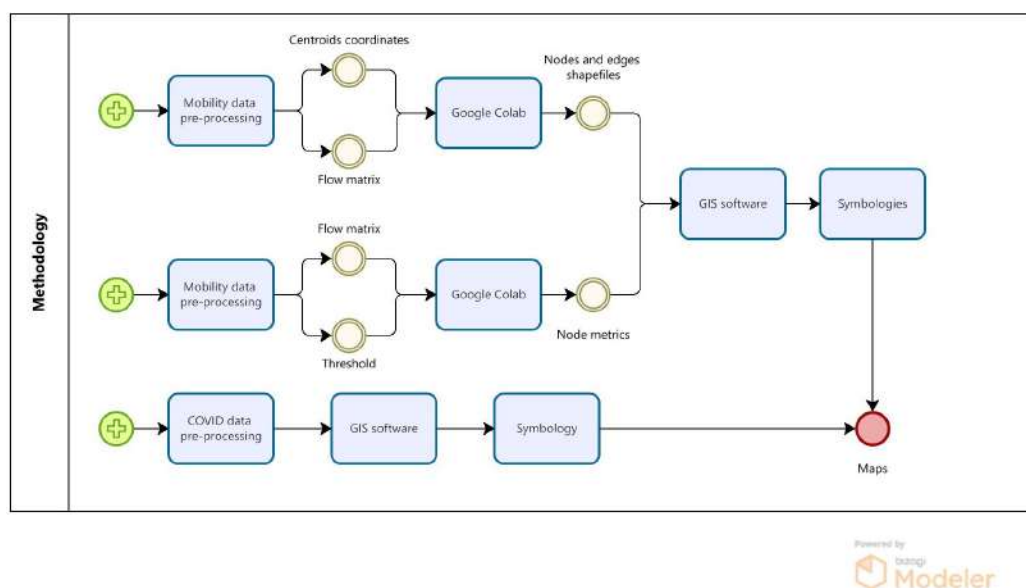


Figura 1. Flowchart of the applied methodology

3. Results

The municipalities of Taubaté and São José dos Campos were the first municipalities to register a case of COVID-19 in the region, in the epidemiological week 12, as can be seen from the representation in Figure 2, followed by the municipalities gradually represented in dark orange and orange. medium. The municipalities of Lagoinha, Queluz, Areias and São José do Barreiro registered their first case between 15 and 19 weeks after the first case registered in the region.

For the distribution of data in classes in the symbology, natural breaks were used. The lowest values of the indexes used are represented in green, with the colors gradually varying up to the highest values, which are represented in red. To represent the occurrence of the first case of COVID-19, the same color scale was used, but oppositely, representing the municipalities that had the first case registered before in red, gradually until the registration of the last first case in green.

The color scale of the indicators represents in red which ones are stronger, highlighting the centroids of the municipalities that have the highest indices. The color scale of the representation of the municipalities highlights in red those municipalities that had the first case registered before.

Even with the geographic proximity between the municipalities of Lagoinha and Taubaté, which share the border, the registration of the first case in each of these cities occurred 15 weeks apart, with Lagoinha registering its first case in epidemiological week 30. If the hypothesis that mobility was one of the factors that led to this large time interval between the first occurrence of the municipalities, since only the physical connection (roads) is not enough for the scattering to occur, requiring a flow of contaminated people.

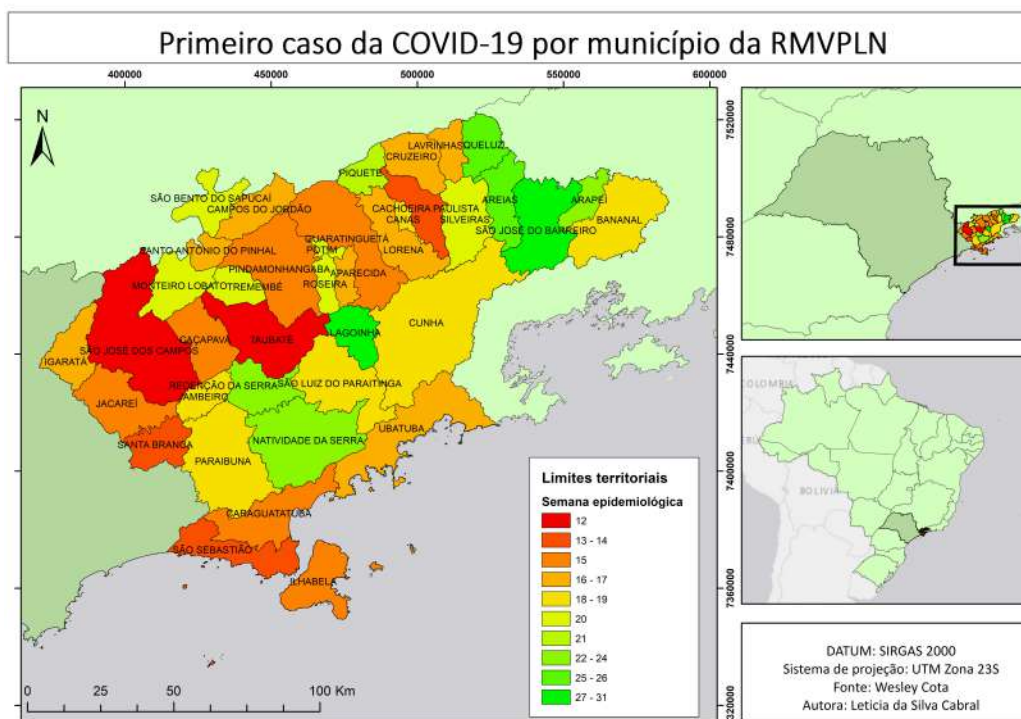


Figura 2. occurrence of the first case of COVID-19 in each municipality of the RMVPLN according to the epidemiological week

With the union of the mobility network and the map with the first case of each municipality, we correlated the occurrences and the calculated topological indices, as well as the flow between the municipalities. In Figure 3, it is possible to notice that the municipalities with many connections and very strong connections. The municipalities of São José dos Campos and Taubaté, which were the first municipalities in the region to register

the first case, have the highest levels of degree centrality and network intermediation. In addition, the largest flow occurs between SJC and Jacareí, followed by the number of trips between the municipality of Taubaté and Tremembé, Pindamonhangaba and São José dos Campos.

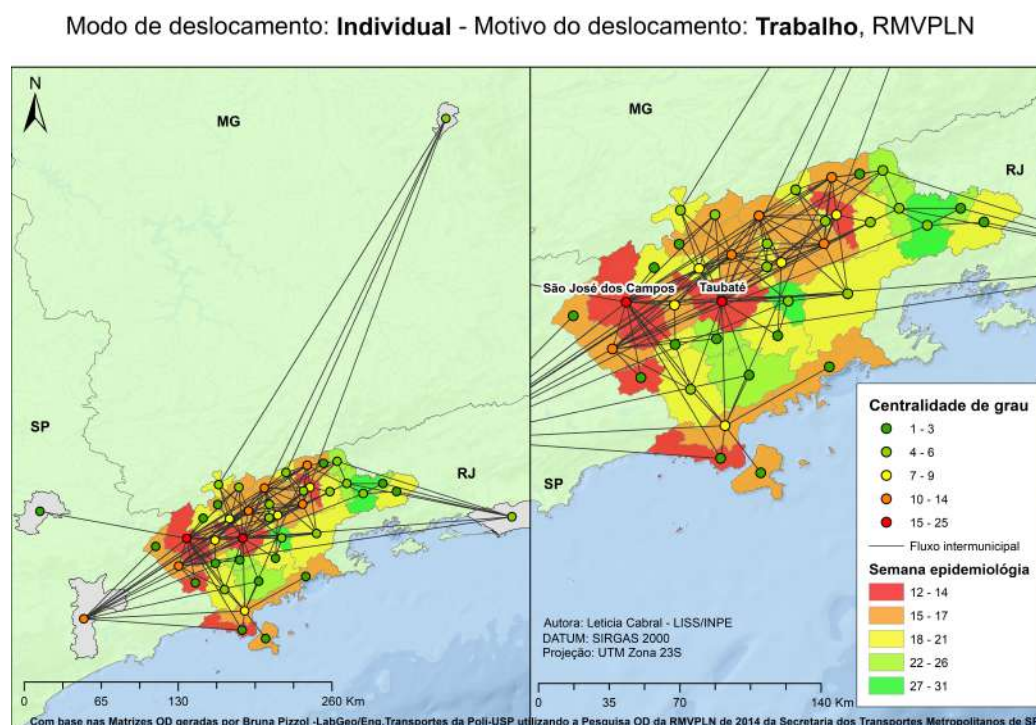


Figura 3. (geo)graph of mobility by individual travel mode for the work reason of the RMVPLN

As we can see, the municipalities with the highest mobility rates evaluated in this study were the ones that first recorded their first occurrence of COVID-19.

4. Conclusion

Given the current scenario, it is important to assess, quantify and analyze mobility and how it occurs, to relate it to the spread of contagious viruses that cause diseases such as SARS-COV-II. The use of the developed methodologies has shown to be promising for carrying out such analyses since it was possible to evaluate the mobility and transport networks in the region and relate them to the occurrences of the first case of COVID-19. In the next steps of the research, we aim to expand the analysis of the transport network, evaluating its impact in the presented context and how its presence and use can result in different scenarios of the spread of COVID-19 in Brazil.

Referências

BOAVENTURA Netto, O. P. (2012). *Grafos: Teoria, modelos, algoritmos*. Editora Bluscher, São Paulo, 5th edition.

- EMPLASA (2016). Região metropolitana do vale do paraíba e litoral norte. <https://emplasa.sp.gov.br/RMVPLN>.
- FORTALEZA, C. M. C. B.; GUIMARÃES, R. B. C. R. C. F. C. P. A. G. B. P. E. (2020). Elementary spatial structures and dispersion of covid-19: health geography directing responses to public health emergency in são paulo state, brazil. *medRxiv*.
- FREITAS, V. L. S., M. G. J. P. S. L. B. L. (2020a). Análise de robustez em uma rede de mobilidade intermunicipal: modelando iniciativas municipais, estaduais e federais como falhas e ataques à contenção do sars-cov-2. *PeerJ*.
- FREITAS, V. L. S.; KONSTANTYNER, T. C. R. O. F. J. S. C. S. N. S. L. B. L. (2020b). The correspondence between the structure of the terrestrial mobility network and the emergence of covid-19 in brazil. *Cadernos de Saúde Pública*.
- IBGE (2020). Campinas/sp, florianópolis/sc e vitória/es passam a estar entre as 15 metrópoles do país. <https://agenciadenoticias.ibge.gov.br/agencia-sala-de-imprensa/2013-agencia-denoticias/releases/28042-regic-2018-campinas-sp-florianopolis>
- JARDIM, A. P. *Reflexões sobre os deslocamentos populacionais no Brasil*. IBGE, Rio de Janeiro, 1st ed edition.
- PERES, R. B. *Covid-19 - crises entremeadas no contexto de pandemia (antecedentes, cenários e recomendações)*. Comissão Permanente de Publicações Oficiais e Institucionais - CPOI, São Carlos, 1st ed edition.
- Sanarmed (2021). Linha do tempo do coronavírus no brasil. <https://www.sanarmed.com/linha-do-tempo-do-coronavirus-no-brasil>.
- SANTOS, L. B. L. e. a. (2017). (geo) graphs-complex networks as a shapefile of nodes and a shapefile of edges for different applications. *arXiv preprint*.

Um Método de Verificação de Conformidade do Plantio Realizado no Contexto do Seguro Rural no Brasil com Zoneamento Agrícola de Risco Climático (ZARC)

Webert S. Pinto¹, Marconi A. Pereira², José R. A. Júnior³, Filipe Santos¹, Gustavo E. Marcatti⁴, Pedro A.A. Silva¹, André Hirsch¹, Andréa C. R. Santos¹, Ivair Gomes³, Fernando M. Pimenta¹, Glauciana M. Ataíde⁴, Gabriel Pereira⁴, Renato V. O. Castro⁴, Marcela C. Ferreira⁴, Leonardo B. Prado³, Fernando A. Teixeira², Daniel C. Pinheiro³

¹ Departamento de Ciências Agrárias – Universidade Federal de São João Del Rei (UFSJ)
Caixa Postal 36307-352– Sete Lagoas – MG – Brasil

² Departamento de Tecnologias em Engenharia Civil, Computação e Humanidades - Universidade Federal de São João Del Rei (UFSJ) Caixa Postal: 36490-972 – Ouro Branco – MG – Brasil

³ Departamento de Geociências - Universidade Federal de São João Del Rei (UFSJ) Caixa Postal: 35701-970 – São João del Rei– MG – Brasil

⁴ Departamento de Engenharia Florestal - Universidade Federal de São João Del Rei (UFSJ) Caixa Postal: 35701-970 – São João del Rei– MG – Brasil

webert.saturnino@gmail.com, marconi@ufsj.edu.br, almeidajr.vr@gmail.com, filipesantos.1f@gmail.com, gustavomarcatti@ufsj.edu.br, pedro@ufsj.edu.br, hirsch_andre@ufsj.edu.br, andreacr2010@gmail.com, ivair@ufsj.edu.br, fernando.m.pimenta@gmail.com, glauciana@ufsj.edu.br, pereira@ufsj.edu.br, renatocastro@ufsj.edu.br, mcferreira.edu@gmail.com, leonardoprado@gmail.com, teixeira@ufsj.edu.br, dcalbino@ufsj.edu.br

Abstract. *The Agricultural Zoning of Climate Risk (ZARC) is an agricultural policy, which identifies regions and moments of lower climate risk for planting crops. The Rural Insurance Subsidy program (PSR) is an important instrument of Brazilian agricultural policy, as it grants insurance subsidies to rural properties. To be entitled to the federal subsidy for the rural insurance premium, the rural producer must respect the sowing or planting data recommended by the ZARC for the municipality. The present work of agricultural cultivation affirms information about the data of rural cultivation and the ZARC, a safer mode of production for the production of crops and guarantees agricultural production, with less.*

Resumo. *O Zoneamento Agrícola de Risco Climático (ZARC) é uma plataforma de política agrícola que define regiões e épocas de menor risco climático para plantio das culturas. O programa de Subvenção ao Seguro Rural (PSR) é um instrumento que concede subvenção ao Prêmio de Seguro Rural para propriedades rurais. Para fazer jus à subvenção, o produtor deve respeitar as datas de semeadura ou plantio preconizadas pelo ZARC. O presente trabalho propõe um algoritmo para cruzamento entre as bases de dados do seguro rural e da plataforma ZARC, de modo a fornecer informações sobre os períodos mais favoráveis para plantio das culturas agrícolas e garantir uma produção agrícola mais assertiva, com menos perdas.*

1. Introdução

O Zoneamento Agrícola de Risco Climático - ZARC (Brasil, 2003) é um instrumento de política agrícola e gestão de riscos na agricultura, que teve início na safra de 1996. O ZARC foi elaborado para reduzir riscos em plantios relacionados a fenômenos climáticos, além de proporcionar para cada município uma estimativa do melhor período para o cultivo de determinada safra (STEINMETZ; SILVA, 2017).

No ZARC são utilizados dados referentes ao clima, ao solo e às plantas, além da aplicação de funções matemáticas e estatísticas para quantificar o risco de perda de lavouras devido a eventos climáticos adversos (SANTOS; MARTINS, 2016; BRASIL, 2021).

Os primeiros zoneamentos agrícolas surgiram há pelo menos quatro décadas, com o objetivo de identificar regiões com menor e maior risco climático para o plantio de determinadas culturas. Nas décadas seguintes, ao perceber o potencial de uso dos zoneamentos, o governo brasileiro passou a adotar o ZARC como ferramenta prática para definição das áreas aptas ou não às diferentes culturas agrícolas produzidas no território brasileiro.

O Programa Nacional de Zoneamento Agrícola de Risco Climático, é regido pelo Decreto Nº 9.841/2019, e permite melhorar a qualidade e disponibilidade dos dados agroclimáticos no Brasil. Com isso, nos diferentes tipos de solo e ciclos de cultivares é possível identificar a melhor época de plantio ao longo do ano, permitindo uma safra mais saudável e de menor risco de perda climática.

Os parâmetros adotados pelo sistema são coletados de uma metodologia fornecida pela Empresa Brasileira de Pesquisa Agropecuária (EMBRAPA), sendo o resultado do estudo é publicado por meio de Portarias da Secretaria de Política Agrícola do MAPA.

O Governo Federal através do Programa de Subvenção ao Prêmio do Seguro Rural (PSR) oferece ao agricultor a oportunidade de segurar sua produção com custo reduzido, subsidiando parte do prêmio que o produtor necessita pagar para acessar os produtos do seguro. Para fazer jus à subvenção federal ao prêmio do seguro rural, o produtor rural deve respeitar as datas de semeadura ou plantio preconizadas pelo ZARC para o município e estar adimplente com o Cadin - Cadastro informativo de créditos não quitados do setor público federal.

Dessa forma, o governo pode contribuir para que o produtor rural continue exercendo as atividades agropecuárias e ao mesmo tempo, abasteçam o setor primário brasileiro e contribuam para o crescimento do Produto Interno Bruto (PIB) do país. Os dados das apólices de seguro rural brasileiras são disponibilizados na plataforma Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER), responsável pela operacionalização do PSR, através de troca de informações entre o atlas do Seguro Rural e dados abertos no portal da transparência.

Neste contexto, o objetivo do presente trabalho é propor um algoritmo para realizar o cruzamento entre dados do Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER) e informações climáticas do Zoneamento Agrícola de Risco Climático (ZARC).

2. Material e Métodos

O ZARC tem os seus dados disponíveis em uma plataforma online¹, fornecidos pelo MAPA, onde a tela é apresentada em uma ferramenta de análise de dados online chamada QlikSense. Para calcular os riscos e recomendações de plantio de uma determinada cultura é necessário informar os seguintes dados: Safra, Cultura, UF, Grupo, Solo e Município. O sistema então apresenta a indicação das épocas de plantio e percentual de risco associado para a cultura no local informado.

Os dados do sistema de Subvenção econômica ao Prêmio do Seguro Rural - SISSER que são utilizados para a operacionalização do Prêmio do Seguro Rural são dados abertos disponibilizados na plataforma online do Ministério da Agricultura, Pecuária e Abastecimento - MAPA². Nesta plataforma constam os dados das culturas agrícolas seguradas, tais como local (município e estado) e informações sobre as indenizações e sinistros associados.

Para o cruzamento entre os dados das apólices de seguro rural e a plataforma do ZARC foi proposto um algoritmo, com vistas a comparar os parâmetros do Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER) com as informações climáticas no ZARC. Para cada apólice selecionada, foi realizado um cruzamento entre as informações das apólices do seguro rural e as informações disponibilizadas na plataforma do ZARC, conforme apresentado na Figura 1.

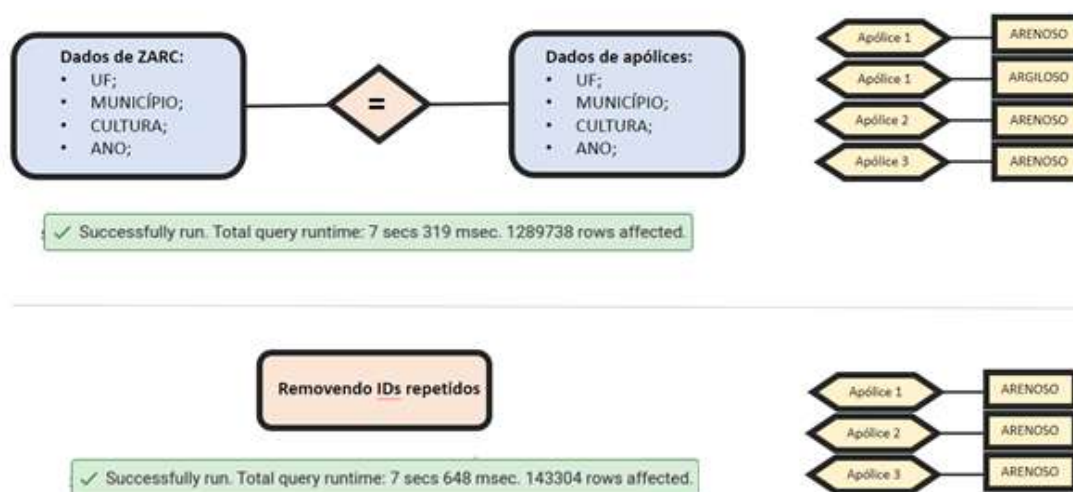


Figura 1. Estrutura do cruzamento dos dados das apólices de seguro rural com a plataforma do Zoneamento Agrícola de Risco Climático (ZARC).

¹ <https://indicadores.agricultura.gov.br/zarc/index.htm>

² <https://dados.gov.br/dataset/sisser3>

Como a plataforma do ZARC solicita a informação do tipo de solo presente na propriedade para cálculo do zoneamento, foi realizada uma parametrização dos solos, haja vista que esta informação não é disponibilizada na base de dados do SISSER. Segundo a Instrução Normativa SPA/Mapa nº 01/2022 (BRASIL, 2022), para o ZARC os solos são agrupados de acordo com a capacidade de retenção de água: a) baixa capacidade – solos tipo I – textura arenosa; b) média capacidade – solos tipo II – textura média; e c) alta capacidade – solos tipo III – textura argilosa.

3. Resultados e discussão

A Figura 2 apresenta como o algoritmo proposto utiliza os parâmetros do Sistema de Subvenção Econômica ao Prêmio do Seguro Rural (SISSER) para obter as informações climáticas no ZARC. Os dados da coluna “SISSER” são utilizados como entrada no sistema ZARC a fim de obter os dados climáticos de cruzamento entre as bases.

Regra	SISSER	ZARC
1	Sigla de cada estado brasileiro (SG_UF_PROPRIEDADE)	UF
2	Cultura (NM_CULTURA_GLOBAL) + ciclo	Cultura
3	Data de início da apólice (DT_INICIO_VIGENCIA)	Safra
4	Nome do município da propriedade (NM_MUNICIPIO_PROPRIEDADE)	Município

Figura 2. Parâmetros do SISSER utilizados como entrada no ZARC

A **regra 1** compara os dados de sigla de Estados brasileiros do Sistema Nacional de ZARC com as siglas dos estados brasileiros do SISSER.

A **regra 2** é a responsável por realizar a identificação da cultura e seu ciclo no ZARC, usando os dados do SISSER. Como este último sistema só possui o nome da cultura e não possui informações sobre o ciclo, adotou-se a premissa de escolher o menor ciclo.

A **regra 3** consiste em identificar qual é o ano da safra que está segurada. Assim, o sistema implementado captura o início da vigência da apólice e o compara com os anos de abrangência do ZARC. Esta regra deverá ser atualizada no futuro, buscando-se como data da safra, a data do plantio estimada (usando-se a análise espectral, a ser implementada em etapas futuras).

A **regra 4** compara os dados de municípios de Estados brasileiros do Sistema Nacional de ZARC com os municípios dos Estados brasileiros do SISSER.

Os tipos de solo das propriedades são gerados na regra 2 da seguinte forma: O arquivo de solos shapefile (.shp), disponibilizado pelo IBGE, apresenta a divisão do solo brasileiro em 319 diferentes texturas específicas, surgindo-se a necessidade de um agrupamento de tais classificações em 3 grandes grupos, sendo eles Arenosa, Média e Argilosa. Dessa forma, as classes de solo IBGE foram adaptadas para o ZARC seguindo a seguinte lógica:

1. Se o solo possuir nomenclatura Argilosa será classificado como Argilosa;
2. Se o solo possuir nomenclatura Média será classificado como Médio;
3. Caso a nomenclatura não seja Argilosa ou Média, o solo será considerado Arenosa;

4. Nos casos onde não há menção a nenhuma das 3 nomenclaturas (Arenosa, Média ou Argilosa) adotou-se a classificação Argilosa;
5. A nomenclatura Siltosa foi associada a Médio;

A Figura 3 apresenta alguns exemplos da aplicação da classificação adotada:

Textura	
Solos IBGE	ZARC
arenosa/arenosa pouco cascalhenta e arenosa	Arenoso
arenosa/argilosa, arenosa/média e média	Argiloso
arenosa/argilosa e muito argilosa	Argiloso
arenosa/argilosa, média/argilosa e arenosa/média	Argiloso
arenosa muito cascalhenta e média muito cascalhenta	Médio
argilosa, média e média	Médio
argilosa/muito arenosa, arenosa e média	Médio
argilosa, muito argilosa e muito argilosa	Argiloso
indiscriminada cascalhenta e indiscriminada	Argiloso

Figura 3. Exemplos de agrupamentos de texturas na classificação de solos do IBGE para entrada no algoritmo de cruzamento entre dados do SISSER e do ZARC.

Um fluxograma apresentando a visão geral do algoritmo de cruzamento entre os dados do ZARC e do SISSER é apresentado na Figura 4:



Figura 4. Visão geral do algoritmo de cruzamento entre os dados do ZARC e do SISSER.

4. Considerações Finais

O Programa de Subvenção ao Seguro Rural é um importante instrumento da política agrícola brasileira, possibilitando ao produtor rural indenizações em caso de perdas, permitindo uma safra mais saudável e de menor risco de perda climática. Para as culturas soja, milho 2ª safra e trigo, no ano de 2020 a região Sul foi a que possuiu maiores números de apólices e indenizações.

O Zoneamento Agrícola de Risco Climático (ZARC) também é uma ferramenta da política agrícola brasileira, adotado nacionalmente, para ajudar o produtor rural a planejar a

safras de forma a evitar perdas na produção. Com o cruzamento dos dados das apólices do seguro rural com os dados do ZARC, os produtores conseguem negociar garantias na tomada de crédito de custeio e, conseqüentemente, reduzir o risco na sua produção.

Dessa forma, o algoritmo proposto possibilita a obtenção de análises da plataforma do ZARC nos diferentes tipos de solo e cultivares, possibilitando a identificação da melhor época de plantio ao longo do ano.

Agradecimentos

À UFSJ pela infraestrutura para a realização da pesquisa e ao MAPA (processo 23122.040239/2021-74) e à Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG) (projeto APQ-00718-21) pelo suporte financeiro.

Referências bibliográficas

Brasil, 2003. Lei no 10.823, de 19 de dezembro de 2003.

BRASIL, Ministério da Agricultura, Pecuária e Abastecimento. **Instrução Normativa no 1, de 21 de Junho de 2022**. Brasília, 2022. Adota, no Zoneamento Agrícola de Risco Climático do Ministério da Agricultura, Pecuária e Abastecimento, as seguintes especificações para solos. Diário Oficial da República Federativa do Brasil, 21 Jun. Acesso em 09/09/2022.

BRAZIL, Ministry of Agriculture Livestock and Food Supply. Adapting to climate change: Strategies for Brazilian agricultural and livestock systems. Brasília, 2021. p. 187.

SANTOS, W. G.; MARTINS, J. I. F. O Zoneamento Agrícola de Risco Climático e sua contribuição à agricultura brasileira. **Revista de Política Agrícola**, v. 25, n. 3, p. 73-94, 2016.

STEINMETZ, S; SILVA, S.C da. Início dos estudos sobre zoneamento agrícola de risco climático (ZARC) no Brasil. **Santo Antônio de Goiás: Embrapa Arroz e Feijão**, 2017.

Padrões espaço-temporais da ocorrência do fogo na Amazônia brasileira

Cândida Caroline S. de S. Leite¹, Liana O. Anderson², Ana Larissa Ribeiro de Freitas², Igor Santiago Broggio³

¹Programa Queimadas – Instituto Nacional de Pesquisas Espaciais (INPE)

Caixa Postal 515 – São José dos Campos – SP – Brasil

²Centro Nacional de Alerta e Monitoramento de Desastres Naturais – São José dos Campos – SP – Brasil

³Universidade Estadual do Norte Fluminense Darcy Ribeiro

CEP: 28013-602 - Campos dos Goytacazes - RJ, Brasil

candida.leite@inpe.br, liana.anderson@cemaden.gov.br,

alarisig@gmail.com, igor@pq.uenf.br

Abstract. *The occurrence of fires in the Amazon can be amplified by severe droughts and the accumulation of combustible material due to deforestation and forest degradation. This study aimed to identify the pattern of occurrence of hotspots between 2003 and 2019, with emphasis on the “fire day”. Hotspot data were used to identify spatio-temporal patterns of hotspot occurrence and positive fire anomalies in the Brazilian Amazon biome. The results showed peaks in the occurrence of hotspots in 2005 and 2007, with respectively 85,721 and 90,206 hotspots observed during the month of September. About 55.3% of the biome area has a monthly recurrence of positive fire anomalies for up to 3 months.*

Resumo. *A ocorrência de incêndios na Amazônia pode ser amplificada por secas severas e o acúmulo de material combustível devido ao desmatamento e à degradação florestal. Este estudo teve como objetivo identificar o padrão de ocorrência de focos de calor entre 2003 e 2019, com destaque para o “dia do fogo”, nesse último ano. Foram utilizados dados de focos de calor para identificar padrões espaço-temporais da ocorrência do fogo e as anomalias positivas desses eventos no bioma amazônico brasileiro. Os resultados demonstraram picos na ocorrência de focos de calor em 2005 e 2007, com respectivamente 85.721 e 90.206 focos observados durante o mês de setembro. Cerca de 55,3% da área do bioma apresenta recorrência mensal de anomalias positivas de fogo de até 3 meses.*

1. Introdução

Na Amazônia, a frequência e a intensidade dos incêndios florestais são reflexo da utilização do fogo no manejo do solo em terrenos de uso agropecuário e das mudanças nos padrões climáticos. As secas ocasionadas por diferentes fenômenos naturais como o El Niño/La Niña e as anomalias de temperatura do Atlântico e do Pacífico que influenciam de forma direta na propagação do fogo e acentuam seus impactos negativos são exemplos dessa dinâmica [ARAGÃO et al., 2018; ANDERSON et al., 2015; SILVA JUNIOR et al., 2019].

As secas também contribuem para o escape do fogo em áreas de florestas, como o que aconteceu no El Niño de 1997/98, no qual cerca de 550 milhões de hectares da Amazônia brasileira se tornaram inflamáveis, sendo que 3,9 milhões destes foram de florestas que realmente queimaram, liberando imediatamente cerca de 200 a 600 Tg de carbono na atmosfera [COPERTINO et al., 2019].

Nesta pesquisa, buscou-se, a partir do recorte temporal de 2003 a 2019, identificar o padrão de ocorrência de focos de calor, com destaque para o dia do fogo, evento antrópico que aconteceu em 10 agosto de 2019 e as anomalias positivas de focos de calor relacionadas.

2. Metodologia

2.1. Área de estudo

O bioma amazônico, constituído pela maior floresta tropical do mundo, ocupa uma extensão de 4.198.999,182 km² do território nacional [MMA, 2019]. Essa área é coberta por 81,2% de Formação Florestal e 14,3% estão direcionadas à Terras Produtivas (Figura 1).

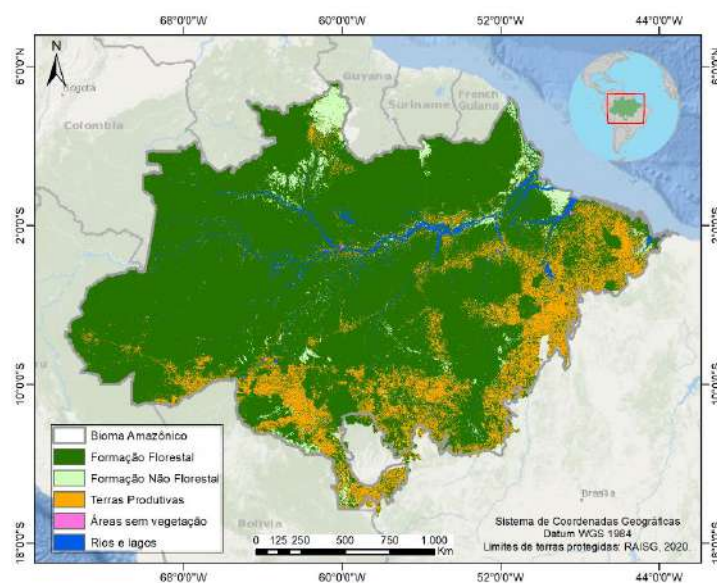


Figura 1. Mapa de localização da área de estudo.

2.2. Aquisição e processamento de dados

Os dados de focos de calor detectados no bioma amazônico foram adquiridos para o período de 2003 a 2019 por meio do Programa Queimadas [INPE, 2019]. Os dados foram coletados em formato vetorial do satélite AQUA_M-T, referência do Programa. Esse dado consiste em pontos de resolução espacial de 1km, com informações sobre a intensidade do fogo, data da observação (dia, mês e ano) e localização. Foram utilizados para a identificação do padrão espaço-temporal de ocorrência da queima.

Durante o processamento, foram extraídas a contagem de focos dentro de uma grade com resolução de 10x10 km, seguida pela conversão para formato matricial. Após esse procedimento foi realizado o cálculo da anomalia de fogo, com 99% de significância, baseado na Equação (1).

$$Anomalias_{\text{focos de calor}}(mês,i) = \frac{Focos_{(mês,i)} - média\ de\ focos_{(mês,i)}}{desvio\ padrão\ de\ focos_{(mês,i)}} \quad (1)$$

Foram consideradas anomalias significativas aquelas maiores ou iguais a 1,99 desvio padrão, o que corresponde a um nível de significância de 99% [Anderson et al., 2010]. O intervalo de confiança foi estabelecido da seguinte forma conforme, $P \leq 0,01$, ou seja, escore $z \geq 2,57$ [Lee e Wong, 2001].

Os procedimentos realizados no presente estudo podem ser divididos em duas etapas: (1) a identificação da causa dos eventos de queimadas e incêndios; (2) o padrão de ocorrência desses. O fluxograma na Figura 2 apresenta a metodologia.

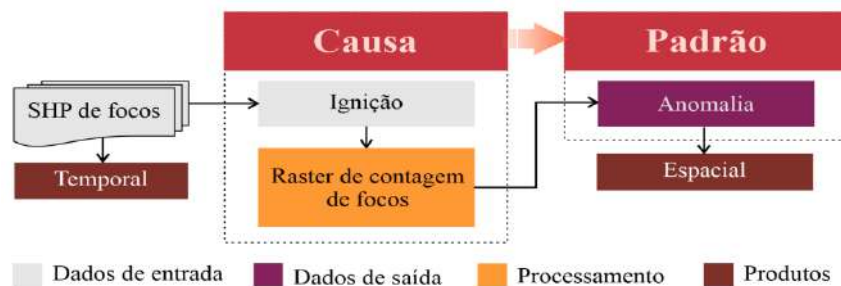


Figura 2. Fluxograma metodológico.

3. Resultados e Discussões

Ao analisar a dinâmica dos eventos de fogo na área de estudo (Figura 3a), destacam-se os meses de agosto e setembro. Em 2005, 2007 e 2010 observam-se picos com 79.387, 55.911 e 56.220 focos de calor em agosto, enquanto setembro apresenta os maiores valores em 2004, 2005, 2006 e 2007 com 82.852, 85.721, 62.055 e 90.206 focos, respectivamente (Figura 3a). O mês de setembro demonstra a maior incidência de focos em todos os anos. No entanto, com o acontecimento do chamado “dia do fogo” (10 de agosto de 2019), ocasião na qual produtores agrícolas realizaram queimadas criminosas de maneira organizada [SILVEIRA et al., 2020], o presente estudo detalhou o mês de agosto, a fim de identificar o dia exato através dos focos de calor.

Observa-se que para o ano de 2019 os dias mais críticos em agosto foram 11, 13 e 14, com 2.706, 2.549 e 2.909 focos de calor nessa mesma ordem (Figura 3b). A figura também chama atenção para o ano de 2005, o qual teve mais de 1.500 focos de calor todos os dias do mês e anos como 2009, 2011 e 2013, anos com menores quantidades de focos, com médias de 5.834, 5.506, 5.298 por ano.

O impacto econômico dos incêndios no bioma Amazônia foi estudado por Campanharo et al. (2019) os quais estimaram uma perda média de $16 \pm 3,42$ milhões de dólares entre 2008 e 2012 e um total de $243,36 \pm 85,05$ milhões de dólares em 2010 pelos prejuízos causados pelas queimadas e incêndios florestais no estado do Acre.

Observa-se na Figura 3 que o padrão temporal de queimadas indica uma interação com os eventos de seca, os quais foram severos nos anos de 2005, 2010 e 2015 [SILVEIRA et al., 2020]. Aragão et al. (2007), Aragão et al (2018) demonstraram a sinergia entre os eventos de secas extremas com os grandes incêndios florestais, mostrando que os dois

processos combinados estão relacionados à ocorrência de anomalias positivas de fogo (APF) na Amazônia.

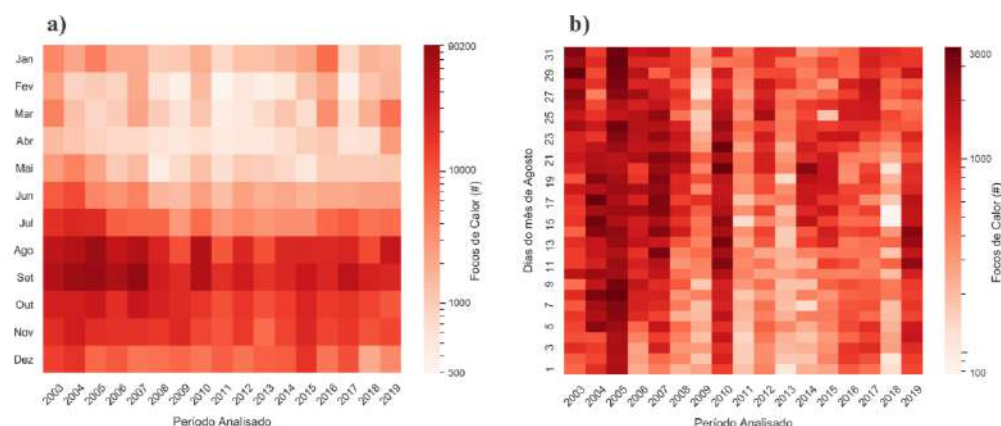


Figura 3. Padrão temporal da ocorrência de focos de calor para o período analisado no bioma amazônico. Em a) Focos de calor para todos os meses da série e b) Focos de calor do mês de agosto.

Na Figura 4 é possível observar a ocorrência de células com APFs significativas para o bioma Amazônico. Além disso, é possível observar por quantos meses uma célula apresentou APF durante a série temporal estudada. As células que foram anômalas por mais de 6 meses somaram uma extensão de 395.400 km² e se concentraram no estado de Roraima e na porção sul do bioma Amazônia, principalmente no estado de Mato Grosso e na faixa que se estende na porção leste, se estendendo desde o Maranhão até o estado de Rondônia, dentro da região denominada de Arco do Desmatamento [COHEN et al., 2007]. Esta região é destacada por Silva Júnior et al. (2019) como áreas que apresentam fortes anomalias negativas de chuva e positivas de temperatura.

Além disso, uma área equivalente a 1,236 milhões de km² apresentou exposição ao fogo, essa área corresponde ao número de células (12.361 células), que apresentaram APF entre 1 e 3 meses ao longo da série estudada. Ou seja, apenas cerca de 55,3% da Amazônia teve 3 meses ou menos de anomalias positivas de fogo.

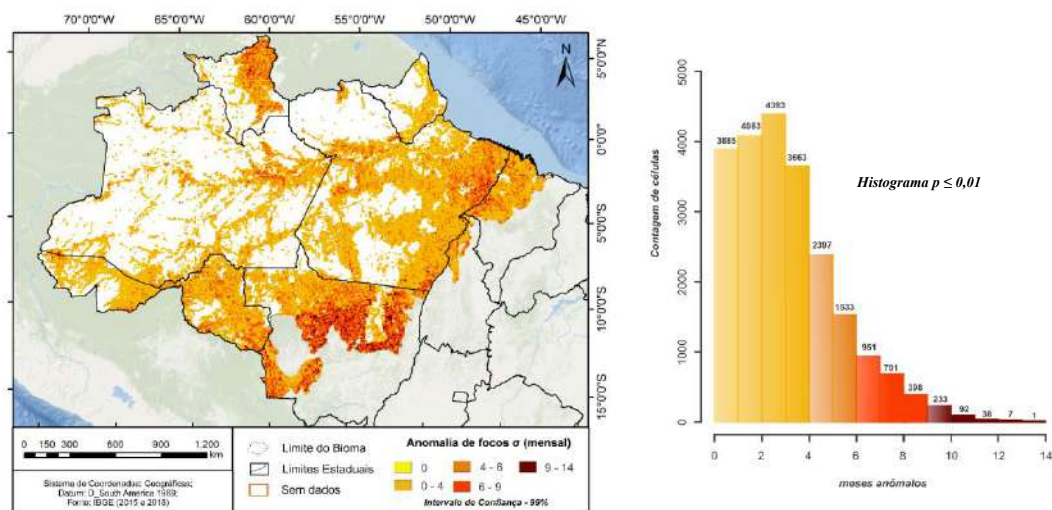


Figura 4. Distribuição espacial da ocorrência de anomalias positivas de focos de calor (APF) para o bioma Amazônico entre 2003 e 2019 e ao lado seu respectivo histograma. Essa análise levou em consideração apenas anomalias positivas de focos de calor dentro do intervalo de confiança de 99% ($p \leq 0,01$).

As APF observadas na Figura 4 também estão associadas a múltiplas fontes de ignição. Nos estados do Amazonas, Pará e Mato Grosso, por exemplo, os incêndios estão relacionados ao manejo da terra para as atividades agropecuárias [WITHEY et al., 2018]. Por outro lado, as ocorrências de incêndios no estado do Maranhão são desencadeadas em áreas de extração seletiva de madeira, de forma criminosa [SILVA JÚNIOR et al., 2019]. Enquanto no estado de Roraima, o uso do fogo é dominado pela prática de subsistência para o manejo na agricultura [FONSECA, 2016].

7. Considerações finais

Os resultados deste estudo permitiram identificar e estimar a área atingida por fogo e o padrão temporal da ocorrência de fogo no bioma Amazônia brasileiro. Além disso, observou-se que no dia do fogo, 10 de agosto de 2019, ocorreu alteração no padrão histórico da ocorrência de queimadas e incêndios florestais na Amazônia. Portanto, há indicativos de que houve uma ação criminosa capaz de aumentar a ocorrência do fogo de forma articulada.

Os dados de focos de calor auxiliaram na identificação dos locais mais afetados por anomalias positivas dessa variável no período estudado e a alta recorrência delas na região do Arco do Desmatamento e no estado de Roraima. A distribuição de quantidade de células anômalas positivas nos meses dos anos estudados também apresentou padrão bastante marcado, separando as ocorrências em baixa quantidade de células anômalas de novembro a maio, enquanto as altas quantidades são percebidas entre junho e outubro.

Referências Bibliográficas

- Anderson, L. O., Malhi, Y., Aragão, L. E., Ladle, R., Arai, E., Barbier, N., et al. Remote sensing detection of droughts in Amazonian forest canopies. *New Phytol.* 187, 733–750. doi: 10.1111/j.1469-8137.2010.03355. 2010.
- Anderson, L. O.; Aragão, L. E. O. C.; Gloor, M.; Arai, E.; Adami, M.; Saatchi, S. S.; Malhi, Y.; Shimabukuro, Y. E.; Barlow, J.; Berenguer, E.; Duarte, V. Disentangling the contribution of multiple land covers to fire-mediated carbon emissions in Amazonia during the 2010 drought. **Global Biogeochemical Cycles**, v. 29, n. 10, p. 1739–1753, 1 out. 2015.
- Aragão, L. E. O. C.; Anderson, L. O.; Fonseca, M. G.; Rosan, T. M.; Vedovato, L. B.; Wagner, F. H.; Silva, C. V. J.; Silva Junior, C. H. L.; Arai, E.; Aguiar, A. P.; Barlow, J.; Berenguer, E.; Deeter, M. N.; Domingues, L. G.; Gatti, L.; Gloor, M.; Malhi, Y.; Marengo, J. A.; Miller, J. B.; Phillips, O. L.; Saatchi, S. 21st Century drought-related fires counteract the decline of Amazon deforestation carbon emissions. **Nature Communications**, v. 9, n. 536, p. 12, 13 dez. 2018.
- Campanharo, W. A.; Lopes, A. P.; Anderson, L. O.; Da Silva, T. F. M. R.; Aragão, L. E. O. C. Translating fire impacts in Southwestern Amazonia into economic costs. **Remote Sensing**, v. 11, n. 7, 2019.
- Cohen, J. C. P.; Beltrão, J. C. B.; Gandu, A. W.; Silva, R. R. Influência do desmatamento

- sobre o ciclo hidrológico na Amazônia. **Ciência e Cultura**, v. 59, n. 3, 2007.
- Copertino, M.; Piedade, M. T. F.; Vieira, I. C. G.; Bustamante, M. Desmatamento, fogo e clima estão intimamente conectados na Amazônia. **Tendências**, v. 71, n. 4, p. 5, out. 2019.
- Fonseca, M. G. Modelling fire probability in the Brazilian Amazon using the maximum entropy method. p. 955–969, 2016.
- Lee J, Wong DWS. *Statistical analysis with ArcView GIS*. Hoboken, NJ, USA: John Wiley and Sons, Inc. 2001.
- Silva Junior, C. H. L.; Anderson, L. O.; Silva, A. L.; Almeida, C. T.; Dalagnol, R.; Pletsch, M. A. J. S.; Penha, T. V.; Paloschi, R. A.; Aragão, L. E. O. C. Fire responses to the 2010 and 2015/2016 Amazonian droughts. **Frontiers in Earth Science**, v. 7, n. June, 2019.
- Silveira, M. V. F.; Petri, C. A.; Broggio, I. S.; Chagas, G. O.; Macul, M. S.; Leite, C. C. S. S.; Ferrari, E. M. M.; Amim, C. G. V.; Freitas, A. L. R.; Motta, A. Z. V.; Carvalho, L. M. E.; Silva Junior, C. H. L.; Anderson, L. O.; Aragão, L. E. O. C. Drivers of Fire Anomalies in the Brazilian Amazon: Lessons Learned from the 2019 Fire Crisis. **Land**, v. 9, n. 12, p. 516, 14 dez. 2020.
- Withey, K.; Berenguer, E.; Palmeira, A. F.; Espírito-Santo, F. D. B.; Lennox, G. D.; Silva, C. V. J.; Aragão, L. E. O. C.; Ferreira, J.; França, F.; Malhi, Y.; Rossi, L. C.; Barlow, J. Quantifying immediate carbon emissions from El Niño-mediated wildfires in humid tropical forests. **Philosophical Transactions of the Royal Society B: Biological Sciences**, v. 373, n. 1760, 2018.

Rain-related Tweets in flooding days: a brief analysis on Tamanduateí river basin

Vitor Y. Hossaki¹, Rogério G. Negri¹, Leonardo B. L. Santos², Livia Tomás²,
Jeferson Feitosa¹

¹Sciences and Technology Institute, São Paulo State University (UNESP)
São José dos Campos, 12245-00, Brazil

²Center for Monitoring and Early Warning of Natural Disasters (CEMADEN)
São José dos Campos, 12245-230, Brazil

{vitor.yuichi, rogerio.negri, jefferson.feitosa}@unesp.br

{santoslbl, liviatomas}@gmail.com

Abstract. *Flooding has become an increasingly frequent phenomenon and reason for several economic and human losses worldwide, such as the São Paulo city, Brazil. The literature has shown that Voluntary Geographic Information (VGI) data is a valuable complementary source for monitoring and analyzing hydrological and meteorological events. Face with this motivation, we extracted VGI from the Twitter social network and analyzed the frequency of meteorological-related posts before flooding/non-flooding events and its correlation with physical-meteorological measures. The obtained results indicate that Twitter is a valuable VGI source for monitoring flooding events.*

1. Introduction

The *Volunteered Geographic Information* (VGI) systems, especially online social networks like Twitter, are increasingly incorporated into several disaster monitoring systems as a low-cost alternative to collecting geographic information. Despite not comprising the original purpose, people posting on Twitter their opinions and perceptions about the environment around them are unreasonably acting as an “social sensor” for natural disasters [Hernandez-Suarez et al. 2019]. Though the geolocation information, tweets can be applied for spatial and temporal modeling of recurrent events in urban zones such as floods [Karami et al. 2020].

Although the numerous advantages of the VGI systems, the exponential growth of volume, velocity, and variety of data in the last few years, increased the challenges to perform accurate filtering and analysis of the data. Furthermore, the information on Twitter can carry a metaphorical way of expressing something. There is a social context associated to complex underlying mechanisms in tweets that are not well known to some technologies [Haworth and Bruce 2015].

Beyond it, the integration of VGI data and *in situ* wireless sensors, combined with a software architecture that processes heterogeneous data, calculates statistics, and allows users to visualize through dashboards, have shown positive results to support monitoring systems and decision-making in emergencies, for adverse events such as floods and landslides [Horita et al. 2015].

In [Hossaki et al. 2021] has shown statistical evidence that tweets that contain keywords associated with meteorological context can be used as a tool for rain detection. In this context, this work aims is to analyze the statistical relationship of Twitter data distribution of the keywords applied between flood days and non-flood days.

2. Materials and method

This study aims to analyze the relation between VGI extracted from the Twitter social network with flooding events and physical-meteorological measures. As study area (Section 2.1), it is considered a portion of São Paulo city, Brazil. The methodological steps and the adopted statistical measures are described in Sections 2.2 and 3, respectively. Python 3.10 programming language was adopted in the following discussions. Pandas, Matplotlib, and SciPy was used as ancillary function libraries.

2.1. Study Area

Beyond its fast population growth, São Paulo city represents the greater financial center in South America. It is densely inhabited, with almost 12.4 million people occupying an area of 1521 km² (8,152 inh./km²). It was estimated in 2010 that 674,329 people in São Paulo were living in areas vulnerable to floods and landslides [Fundação Instituto Brasileiro de Geografia e Estatística - IBGE 2022, Hossaki et al. 2022]. Floods and intense rain are increasingly phenomena in the São Paulo region. Population growth allied to disordered urban expansion and the increase of meteorological events in the last years has drawn the attention of researchers about hydrological extremes. [Carvalho 2021].

The study area comprises the Tamanduateí river basin, with 323 km² of area, it extends to the hydrographic basins of Rio Pinheiro, Rio Guaió, Rio Aricanduva and Córrego de Tapuapé. This basin's catchment area belongs to six municipalities, where 115 km² are within São Paulo city. The geolocalized Twitter data and the floods were delimited in a spatial radius of 2000 m and correspond to the localization of the rain gauge 355030857A from the National Center for Monitoring and Early Warning of Natural Disasters (CE-MADEN) (Figure 1).

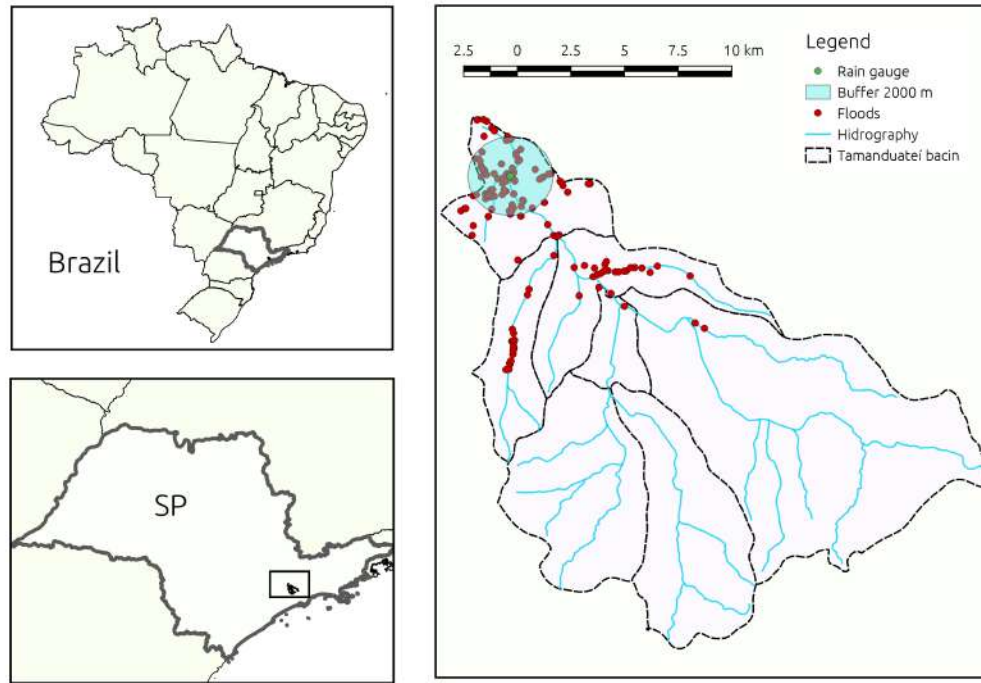


Figure 1. Study Area location at Tamanduatei basin river

2.2. VGI and floods

The VGI data were extracted using the Twitter’s API (*Application Programming Interface*). This data were collected in a spatial radius of 2000 m regarding the location of the CEMADEN automatic rain gauge 355030857A. The period of analysis comprises the first trimester of 2019.

Posterior, a “text-content filtering process” was applied on the Twitter data by considering a list of keywords associated with rain and floods. The keywords were: “chuva”, “chove”, “chuvoso”, “chuvosa” e “chubarada”. According to [de Andrade et al. 2021], these keywords are less spatially and temporally volatile. For further analysis, After that, the tweets were separated into two temporal series expressing the number of keywords detected in every filtered tweet post.

Flood registries were accessed at Emergency Management Center website [São Paulo’s Emergency Management Center - CGE 2022]. These records contains the starting and ending time of blocking and the street name with the road intersection. The vector data with points representing traffic blocks caused by floods, registered from January to March 2019 and inside the 2000 m neighborhood radius of 355030857A rain gauge, were obtained from [Tomás et al. 2022].

3. Statistical Foundation

Hypothesis tests are capable to prove divergences and statistical relationships. These tests $\mathcal{H}$ are constituted by a null hypothesis $\mathcal{H}_0$, which implies no statistical difference, and $\mathcal{H}_1$

expresses the opposite statement. Assuming a sample $\{x_i : i = 1, \dots, n\}$ in data space $\mathbf{A}$, through $\mathcal{H}$ identifying a critical region $C_\tau \subset \mathbf{A}$ that defines extreme situations and take to the rejection of $\mathcal{H}_0$ under a probability α of making a wrong decision.

There are a variety of procedures have been proposed in the literature to verify statistical hypotheses, for example, tests for single-sample tests, two paired or independent samples, correlations, concordance, and time series. Furthermore, these tests are categorized as parametric or non-parametric, where the latter does not consider premises about the Gaussian distribution.

The U Mann-Whitney's test is a non-parametric alternative to compare two independent samples, allowing conclude if such samples come from the same population. Assuming the two sets $\{x_1, \dots, x_a\}$ and $\{y_1, \dots, y_b\}$ with $a < b$, the statistic test $U = \min\{U_a, U_b\}$, such that:

$$\begin{aligned} U_a &= ab + \frac{a(a+1)}{2} - R_a \\ U_b &= ab + \frac{b(b+1)}{2} - R_b \end{aligned} \quad (1)$$

where R_a and R_b are the ranking sums linked to the element from the smaller and larger sets. After ordering its elements into a single set, respectively. If $a < 20$, a specific sampling distribution may be considered to define the assigned p -value [Eplett 1979]. Else, when $a \geq 20$, the statistic U tends to a normal distribution with parameters $\mu = \frac{ab}{2}$ and $\sigma = \sqrt{\frac{ab(a+b+1)}{12}}$, and a single-tailed p -value represents the probability of values equal or higher than U .

4. Results and discussion

Initially, we obtained 79896 tweets in the neighborhood and period of analysis. However, after the filtering process based on the keywords mentioned in Section 2.2, only 492 were selected for the statistical analysis, which means that about 0,6% of the total amount of Tweets contain keywords possibly related with rains and floods.

Figure 2 shows that the tweets frequency in flooding days is superior to non-flooding days. Furthermore, it is evident that the keywords frequency on flood days are superior, the interquartiles and median are higher compared to non-flood days.

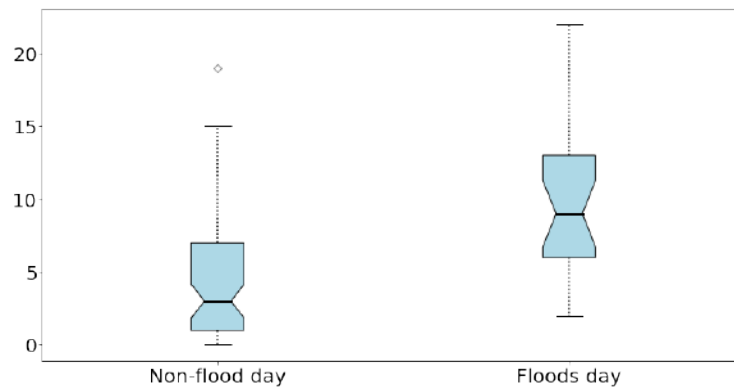


Figure 2. Boxplot of keywords frequency in flooding and non-flooding days.

Additionally, the Twitter's data and floods registry are discrete values that do not follow a specific distribution. In this context, we applied the non-parametric U Mann-Whitney's test to compare the keyword frequencies in flooding and non-flooding days. The statistical results evidence the difference between flood days and non-flood days. The p -value of 6.153×10^{-7} , compared to a significance of $\alpha = 0.05$, allowed to reject the null hypothesis that the samples were drawn from a same population.

Under these conditions, we have that the list of words used in the filtering has greater sensitivity when flooding occurs. Statistical validation indicates that keywords can detect phenomena such as precipitation, as they also appear more frequently on days when the hydrological phenomenon occurs.

5. Conclusion

This work consisted in Data Analysis of geolocated Twitter's information allied with flood registry. We established that, in this study area with the specified time window, there is an important statistical result that words used on flood days may be more relevant than non-flood days. In this context, the work has evidence that keywords not only detect general phenomena but can also lead to greater statistical weighting on days when flooding occurs. Twitter seems to have underlying mechanisms in which words shape the user's perception of a given event. Meantime, there are some challenges in filtering at VGI system related to data quality, it means that the applied keywords can be used to metaphorical posts that are not related to the object of study.

As perspective for future studies, it is pretended to integrate *in situ* and remote sensor with VGI data applying several metrics and data processing methods to figure out ways to development robust models for flood detection and forecasting.

References

- Carvalho, R. B. d. (2021). Análise de chuvas extremas e a relação com eventos de alagamentos na cidade de São Paulo–SP.
- de Andrade, S. C., Porto de Albuquerque, J., Restrepo-Estrada, C., Westerholt, R., Rodriguez, C. A. M., Mendiola, E. M., and Delbem, A. C. B. (2021). The effect of

- intra-urban mobility flows on the spatial heterogeneity of social media activity: investigating the response to rainfall events. *International Journal of Geographical Information Science*, pages 1–26.
- Eplett, W. J. R. (1979). The small sample distribution of a mann-whitney type statistic for circular data. *The Annals of Statistics*, 7(2):446–453.
- Fundação Instituto Brasileiro de Geografia e Estatística - IBGE (2022). Panorama of brazilian municipalities. <https://cidades.ibge.gov.br/brasil/sp/sao-paulo/panorama>. Accessed: 2022-09-18.
- Haworth, B. and Bruce, E. (2015). A review of volunteered geographic information for disaster management. *Geography Compass*, 9(5):237–250.
- Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Perez-Meana, H., Portillo-Portillo, J., Sanchez, V., and García Villalba, L. J. (2019). Using twitter data to monitor natural disaster social dynamics: A recurrent neural network approach with word embeddings and kernel density estimation. *Sensors*, 19(7):1746.
- Horita, F. E., de Albuquerque, J. P., Degrossi, L. C., Mendiondo, E. M., and Ueyama, J. (2015). Development of a spatial decision support system for flood risk management in brazil that combines volunteered geographic information with wireless sensor networks. *Computers & Geosciences*, 80:84–94.
- Hossaki, V., Negri, R., Freitas, C., and Santos, L. (2021). A TWITTER-BASED METEOROLOGICAL RADAR. *INIC XXV*, pages 1–5.
- Hossaki, V., Negri, R., Londe, L., Tomás, L., Santos, R., Seron, W., Feitosa, J., Andrade, S., and Santos, L. (2022). Physical and Social-based rain gauges - a case study on urban floods detection (To be submitted). *Frontiers*.
- Karami, A., Shah, V., Vaezi, R., and Bansal, A. (2020). Twitter speaks: A case of national disaster situational awareness. *Journal of Information Science*, 46(3):313–324.
- São Paulo’s Emergency Management Center - CGE (2022). Flood records in são paulo, brazil in 2019.
- Tomás, L. R., Soares, G. G., Jorge, A. A. S., Mendes, J. F., Freitas, V. L. S., and Santos, L. B. L. (2022). Flood risk map from hydrological and mobility data: A case study in são paulo (brazil). *Transactions in GIS*, 26(5):2341–2365.

A convolutional LSTM neural network for precipitation nowcasting based on weather radar data

Aurelienne A. S. Jorge¹, Marcos Quiles², Izabelly C. Costa¹, Leonardo B. L. Santos³

¹Instituto Nacional de Pesquisas Espaciais (INPE)
Cachoeira Paulista, SP – Brazil

²Universidade Federal de São Paulo (Unifesp)
São José dos Campos, SP – Brazil

³Centro Nacional de Monitoramento e Alertas de Desastres Naturais (Cemaden)
São José dos Campos, SP – Brazil

aurelienne.jorge@inpe.br, quiles@gmail.com, izabelly.costa@inpe.br,
santoslbl@gmail.com

Abstract. *Due to the high instability of the atmosphere, it is a challenging task to forecast precipitation events in the very short term. Artificial Neural Networks have been recently employed in this context as an alternative to solve prediction problems. This work proposes a convolutional LSTM neural network based on weather radar images (spatial data) to forecast rainfall events, whatever the precipitation intensity is. The preliminary results show better performance when comparing the network prediction error with a persistence model.*

1. Introduction

First mentioned by [Browning 1980], nowcasting is defined as the description of the weather's current state and the prediction of changes that can be expected in a few hours - from the present to 6 hours ahead. Due to the high instability of the atmosphere, it is impossible to precisely predict the location and time of convective phenomena several hours in advance. In fact, it is already a challenging task to forecast within a few hours in advance.

Artificial Neural Networks (ANNs) have been designed by researchers in many scientific areas to solve a variety of problems such as pattern recognition, optimization, and prediction, including weather forecasts [Jain et al. 1996]. Kumar et al. [Kumar et al. 2020] propose a precipitation nowcasting architecture named 'ConvCast' that uses a Convolution Long-Short Term Memory (ConvLSTM) neural network to predict short-term precipitation events based on satellite data. Caseri et al. [Caseri et al. 2022] also present a ConvLSTM solution to precipitation nowcasting using weather radar data, albeit focusing on heavy rainfall events.

This work proposes a ConvLSTM neural network based on remote sensing data for precipitation nowcasting. In this initial phase, we use weather radar data, but satellite images may also be included in future research steps. Convolutional neural networks are suitable for data grids like weather radar images as they keep the spatial relations between features [Jorge et al. 2020]. Moreover, LSTM is a recurrent neural network appropriate to perform predictions based on time series data. Therefore, ConvLSTM is an ideal tool to apply in this context of nowcasting using weather radar images.

The focus is not only on heavy rainfall events [Caseri et al. 2022], but also on events of lighter intensity. Initial results show predictions with low error to the observed data. The paper is organized as follows: Section 2 describes the dataset used, the pre-processing flow, and the implemented model architecture. Section 3 presents the preliminary results of this study, and Section 4 contains the final considerations and discusses the future perspectives of this ongoing research.

2. Materials and Methods

The dataset is composed of weather radar images from march 2019. The data was collected with the weather radar located in the city of São Roque (São Paulo - Brazil). This equipment has a spatial coverage of 250 kilometers and provides high spatial and temporal resolution data, with 1 kilometer and 10 minutes. It performs azimuth scans in 360 degrees over 15 different elevation levels composing a volumetric scan. From that, we can extract a product named the Constant Plan Position Indicator (CAPPI), which consists of a horizontal cross-section of data at a constant height. For this work, we use the CAPPI product at the height of 3 km to avoid altitude changing and ground echo problems. The product is a gray-scale image whose pixels represent reflectivity values (in dBz units), from which we can later estimate the precipitation rate. Only values above 20 dBZ were considered in this study, as it is the threshold to represent a light rain.

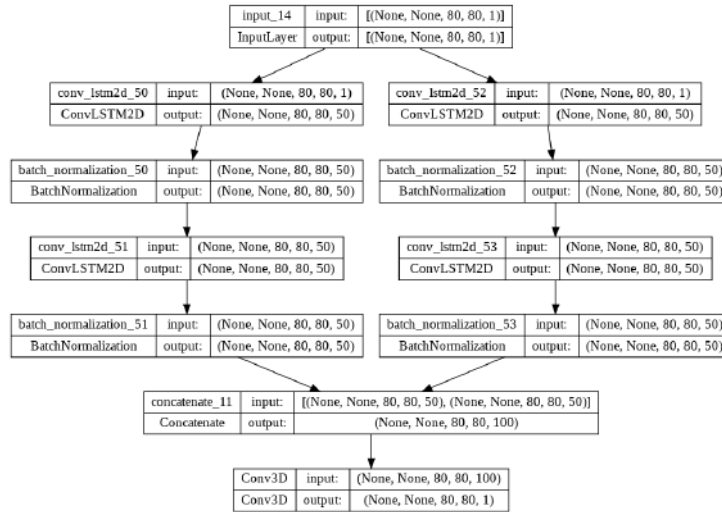


Figure 1. The ConvLSTM topology

In the pre-processing phase, we perform a spatial clipping to select only the area of interest: the Metropolitan Area of São Paulo (MASP). As a result, we have 80x80 grids with 1 kilometer of resolution. In this step, we also split the time series into samples of 10 frames each, keeping the temporal sequence with a time step of 10 minutes. There are 309 samples, which are split into train and test sets sequentially (80% and 20%, respectively). The values are normalized to fit in a range from 0 to 1.

An application written in Python with the Keras library was developed to pre-process the dataset and implement the neural network. The ConvLSTM architecture is presented in Figure 1. The input layer is prepared to receive data with the following

shape: 80x80x1 (width x height x channels). There is no fixed number of time steps to make the model flexible. The topology is mainly composed of ConvLSTM layers, each one followed by a Batch Normalization. The input is received by two branches: the first one is composed of ConvLSTM layers with 50 neurons each and a kernel size of 1x1, and the second one with the same number of layers and neurons but with a kernel size of 3x3. At the end, the two branches are concatenated to feed a 3D convolution layer that delivers the network's output in the same shape as the input data. In the output layer (Conv3D), its size is increased to 1x3x3. The Rectified Linear Activation Function (ReLU) is the activation function of the ConvLSTM layers, and Sigmoid is the activation function of the Conv3D layer. Adam is the optimizer employed to adjust the network weights, and Mean Squared Error (MSE) is the loss function set in the model compilation.

3. Results

The training phase was executed for ten epochs, and 10% of the training set was set apart for the validation process. At the end of the training phase, the result is a validation loss of $1.2e - 2$. Moreover, the loss obtained in the model evaluation, using the test subset, is even lower: $6.8e - 3$. Besides, tests were made with various scenarios to analyze the model performance further.

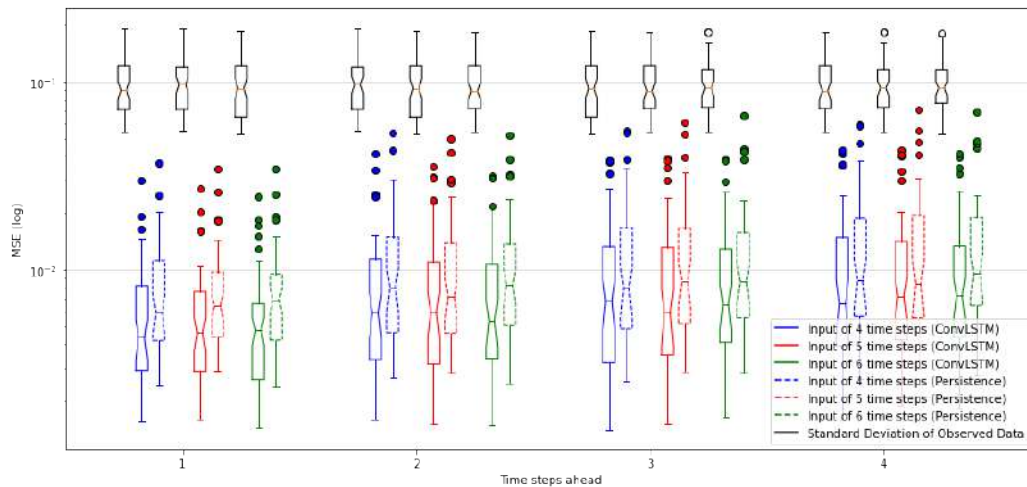


Figure 2. Boxplots of MSE based on 4 (blue), 5 (red) and 6 (green) time steps of observed data as input, and the predictions are for 1, 2, 3 and 4 time steps ahead.

Inputs with 4, 5 and 6 time steps (40, 50 and 60 minutes, respectively) were supplied to the model in order to predict the next 4 time steps (10, 20, 30 and 40 minutes). Figure 2 presents the boxplots with the MSE distribution for each scenario. The results are compared with a persistence model, which consists of keeping the last observed data. Although the range of the distributions is very similar, we notice that the error tends to slightly increase as we try to predict more time steps ahead. There is no significant difference when using more samples of observed data as input. An input of 4 time steps is enough to have a better prediction compared to the persistence model.

Figure 3 illustrates an example of the result generated from the model prediction. It is based on a sample from the test set using 4 time steps (40 minutes of observed data

with 10 minutes step) as input. The predictions are made for the next 4 time steps (10 to 40 minutes ahead).



Figure 3. Example of prediction for 4 time steps ahead (steps 5 to 8 of *Prediction* row) based on 4 initial steps of observed data (steps 1 to 4). The real observed data for steps 5 to 8 are presented on the second row (*Ground Truth*).

In this sample, which shows part of a light rain event, it is possible to observe that the predictions are visually very similar to the ground truth.

4. Final Considerations

This paper presented a convolutional LSTM neural network to forecast rainfall events evolution in a very short time based on weather radar data. Tests were made with different numbers of input time steps to predict precipitation events within 40 minutes ahead. The results show better results when comparing the network prediction error with a persistence model. It could be also observed that there is no significant difference when using more samples of observed data as input. An input of only 4 time steps is enough to have a better prediction compared to the persistence model.

As a continuation of this study, there are perspectives of including more data in the model, such as other physical parameters of the atmosphere, to enable better tracking of event directions and the formation of new events.

5. Acknowledgements

This research was partially supported by grant 420338/2018-7 of the Brazilian National Council for Scientific and Technological Development (CNPq) and by grants 2018/06205-

7 and 420338/2018-7 of São Paulo Research Foundation (FAPESP) and DFG-IRTG 1740/2.

References

- [Browning 1980] Browning, K. A. (1980). *Local weather forecasting*, page 179–211.
- [Caseri et al. 2022] Caseri, A. N., Lima Santos, L. B., and Stephany, S. (2022). A convolutional recurrent neural network for strong convective rainfall nowcasting using weather radar data in southeastern brazil. *Artificial Intelligence in Geosciences*, 3:8–13.
- [Jain et al. 1996] Jain, A., Mao, J., and Mohiuddin, K. (1996). Artificial neural networks: a tutorial. *Computer*, 29(3):31–44.
- [Jorge et al. 2020] Jorge, A. A. S., Costa, I. C., and Santos, L. B. L. (2020). Geographical Complex Networks applied to describe meteorological data. In *Proceedings...*, pages 258–263, São José dos Campos, Brazil. BRAZILIAN SYMPOSIUM ON GEOINFORMATICS, INPE.
- [Kumar et al. 2020] Kumar, A., Islam, T., Sekimoto, Y., Mattmann, C., and Wilson, B. (2020). Convcast: An embedded convolutional lstm based architecture for precipitation nowcasting using satellite data. *PLOS ONE*, 15(3):e0230114.

Daily Weather Pattern for São Paulo city: from 1st January to 31st March, 2019

Giovanni Dolif Neto¹, Eliana Vale¹, Arthur Muller²

¹Centro Nacional de Monitoramento e Alertas de Desastres Naturais
Ministério de Ciência, Tecnologia e Inovações (CEMADEN-MCTI)
Estrada Doutor Altino Bondensan, 500 - Distrito de Eugênio de Melo
São José dos Campos – SP – Brazil

²Rua João Müller Júnior, 175. Picadas do Sul, São José, Santa Catarina, Brazil.

{giovanni.dolif,eliana.vale}@cemaden.gov.br, arthurmuller91@gmail.com

Abstract. *This paper presents the list of weather pattern categories over São Paulo city for every day from 1st January to 31st March. The convective weather pattern is the most frequent (42%), followed by the absence of weather pattern (36%). Although cold fronts, troughs, South Atlantic Convergence Zone and Moisture Convergence Zone weather patterns correspond to a fewer number of days, their impacts in urban mobility can be significant. Therefore, the results of this paper can be useful in future studies correlating, for example, this list of weather pattern categories with urban flash floods. That could contribute to improve flash flood warnings and mitigate their impacts.*

1. General Information

Rain triggered urban flash floods result in considerable impact on urban mobility, causing economic losses and reduction in people welfare during the summer months of January, February and March [Vale 2020, Haddad and Teixeira 2015]. This short paper presents the weather pattern categories for each day from 01st January and 31st March, in order to ease future studies correlating, for example, weather patterns and urban flash floods in São Paulo city.

2. Meteorological Classification Method

The meteorological classification was performed by analysing satellite images [CPTEC 2019a], synoptic charts [MARINHA 2019, CPTEC 2019b] and surface observation [REDEMET 2019]. The following weather pattern categories were considered:

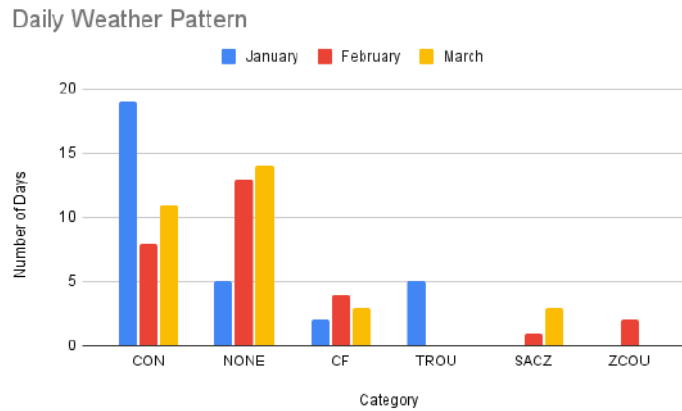
1. cold front (CF)
2. atmospheric mid-level trough (TROU)
3. local convection (CON)
4. Moisture Convergence Zone (ZCOU)
5. South Atlantic Convergence Zone (SACZ)
6. None of the above (NONE)

The ZCOU category was introduced by the weather forecast team of the Brazilian Centre for Weather Forecast and Climate Studies (CPTEC) in last decade [SACRAMENTO NETO et al. 2010, PALLOTTA and Nakazato 2010]. The NONE category means that none of the other categories was identified.

3. Results

The classification method resulted in a predominance of the *CON* category in January (Figure 1). The *NONE* category predominated in February and March. *SACZ* category was not present in January while *CF* was more frequent in February. One episode of *ZCOU* was detected in late February.

Figure 1. Daily meteorological classification and flash flood reports number for January, 2019



The complete list of daily categories can be found in Tables 1, 2, 3

3.1. January

In January, 19 days (62%) were classified as convective (*CON*). *TROU* and *NONE* categories were found in the same 5 (16%) number of days. Only two days were classified as *CF*.

- *CON*: 19 days (62%)
- *NONE*: 05 days (16%)
- *TROU*: 05 days (16%)
- *CF*: 02 days (06%)

3.2. February

In February all the categories were present. The *NONE* category was the most frequent, 46% of the days, followed by *CON* days (29%). Cold fronts were identified in four days, while *ZCOU* or *SACZ* were found in only three days.

- *NONE*: 13 days (46%)
- *CON*: 08 days (29%)
- *CF*: 04 days (14%)
- *ZCOU*: 02 days (07%)
- *SACZ*: 01 days (04%)

3.3. March

In March, no ZCOU episode was found. The NONE category was the most frequent (45%), followed by CON category (35%). The CF and SACZ category were both present in three days.

- NONE: 14 days (45%)
- CON: 11 days (35%)
- CF: 03 days (10%)
- SACZ: 03 days (10%)

4. Conclusion and Discussion

The dominant meteorological weather patterns in each day of January, February and March, 2019 in São Paulo city were identified. The convective (CON) weather pattern was the most frequent (42%), followed by the absence of weather pattern (NONE) category (36%). Trough, SACZ and ZCOU weather patterns corresponded, together, to only 12% of days.

The results presented in this paper could be valuable data for future studies correlating weather pattern with, for example, urban flash floods. The literature reveals that urban flash floods markedly impact urban mobility in São Paulo city, causing significant economic losses [Haddad and Teixeira 2015]. Such a future study could improve flash flood warnings, therefore contributing to mitigate their impacts.

References

- CPTEC (2019a). Goes-16 satellite images. Available at: <http://satellite.cptec.inpe.br/>, Accessed: 01/08/2022.
- CPTEC (2019b). Synoptic charts. Available at: <http://img0.cptec.inpe.br/~rgptimg/Produtos-Pagina/Carta-Sinotica/Analise/Superficie/>, Accessed: 01/08/2022.
- Haddad, E. A. and Teixeira, E. (2015). Economic impacts of natural disasters in megacities: The case of floods in São Paulo, Brazil. *Habitat International*, 45:106–113. Special Issue: Exploratory Spatial Analysis of Urban Habitats.
- MARINHA (2019). Cartas sinóticas.
- PALLOTTA, M. and Nakazato, R. Y. (2010). Caracterização de episódios de zona de convergência do atlântico sul (zcas) e zona de convergência de umidade (zcou) em janeiro e fevereiro de 2010. In *Congresso Brasileiro de Meteorologia*, volume 16.
- REDEMETS (2019). Meteorological terminal aviation routine weather report - metar.
- SACRAMENTO NETO, O. B. d., ESCOBAR, G. C. J., and SILVA, P. E. D. d. (2010). Método objetivo para identificar episódios de zonas de convergência de umidade (zcou) no ambiente operacional do centro de previsão de tempo e estados climáticos-cptec. In *Congresso Brasileiro De Meteorologia*, volume 16.
- Vale, R. C. C. (2020). Quanto custa a imobilidade urbana em São Paulo? *POLICY*, 3.

Table 1. Daily meteorological classification - January, 2019

Date	Category
01/01/2019	CON
02/01/2019	CON
03/01/2019	CON
04/01/2019	CF
05/01/2019	CF
06/01/2019	CON
07/01/2019	NONE
08/01/2019	CON
09/01/2019	NONE
10/01/2019	CON
11/01/2019	CON
12/01/2019	CON
13/01/2019	CON
14/01/2019	CON
15/01/2019	CON
16/01/2019	TROU
17/01/2019	TROU
18/01/2019	CON
19/01/2019	CON
20/01/2019	CON
21/01/2019	NONE
22/01/2019	CON
23/01/2019	CON
24/01/2019	CON
25/01/2019	TROU
26/01/2019	TROU
27/01/2019	TROU
28/01/2019	NONE
29/01/2019	CON
30/01/2019	CON
31/01/2019	NONE

Table 2. Daily meteorological classification - February, 2019

Date	Category
01/02/2019	NONE
02/02/2019	NONE
03/02/2019	CON
04/02/2019	CF
05/02/2019	NONE
06/02/2019	NONE
07/02/2019	SACZ
08/02/2019	NONE
09/02/2019	NONE
10/02/2019	NONE
11/02/2019	NONE
12/02/2019	CON
13/02/2019	CF
14/02/2019	NONE
15/02/2019	CON
16/02/2019	CON
17/02/2019	NONE
18/02/2019	NONE
19/02/2019	NONE
20/02/2019	CON
21/02/2019	CON
22/02/2019	CON
23/02/2019	NONE
24/02/2019	CON
25/02/2019	CF
26/02/2019	CF
27/02/2019	ZCOU
28/02/2019	ZCOU

Table 3. Daily meteorological classification - March, 2019

Date	Category
01/03/2019	SACZ
02/03/2019	SACZ
03/03/2019	SACZ
04/03/2019	CON
05/03/2019	CON
06/03/2019	CON
07/03/2019	CON
08/03/2019	NONE
09/03/2019	CON
10/03/2019	CF
11/03/2019	CF
12/03/2019	CON
13/03/2019	NONE
14/03/2019	CON
15/03/2019	CON
16/03/2019	NONE
17/03/2019	CON
18/03/2019	CON
19/03/2019	CON
20/03/2019	NONE
21/03/2019	CF
22/03/2019	NONE
23/03/2019	NONE
24/03/2019	NONE
25/03/2019	NONE
26/03/2019	NONE
27/03/2019	NONE
28/03/2019	NONE
29/03/2019	NONE
30/03/2019	NONE
31/03/2019	NONE

Development of a data set for distributed hydrological simulation of Paraíba do Sul Basin

Lina M. Osorio Olivos¹, Arisvaldo Vieira Mélo Júnior¹, Silvana Susko Marcellini²,
Dário Hachisu Hossoda², Rogério Moreira Chagas², Gabriel A. dos Santos Soares²,
Flávia Ribeiro dos Santos Jacobsem², André Schardong²

¹Hydraulic and Environmental Engineering Section
Polytechnic School – São Paulo University (USP) – São Paulo, SP – Brazil

²Decision Support System Laboratory (LabSid)
Polytechnic School – São Paulo University (USP) – São Paulo, SP – Brazil

{lmosorioo, arisvaldo}@usp.br, {silvana, dario, rogerio, gabriel,
flavia}@labsid.eng.br, andreschardong@gmail.com

Abstract. *This work describes a dataset created for the purpose of simulating the Paraíba do Sul basin with hydrological distributed models. The dataset contains gridded numerical and descriptive information related to different characteristics of the basin. Distributed hydrological models are useful tools for the management of water resources, as they allow estimating the hydrological regime in a basin according to its physical, rainfall and land use change characteristics, seeking to accurately quantify the availability of water allowing planning future use according to scenarios of land use change or climate change.*

Keywords: Gridded Data, Distributed models, Spatio-Temporal series, Hydrology

1. Introduction

Basic data related to climate, topography, land cover and soil characteristics are necessary for the hydrological simulation of basins [Xavier, et.al. 2016]. Quantifying water balance in tropical areas requires knowledge of the influence of heterogeneity in soil-vegetation-atmospheric relationships [Mélo Junior, et al. 2022]. Distributed models achieve to represent the spatial-temporal variation in the variables that force the hydrological cycle, however the availability of these type of data is a challenge in many countries [Kepeng and Juncang 2021].

Located in the South-East, the Paraíba do Sul (Figure 1) is one of the most significant river basins of Brazil in terms of urban water supply. This interstate basin forms part of the states of Minas Gerais, Rio de Janeiro and São Paulo, and presents particular socio-economics characteristics and a significant increasing in water demand in the last years [OCDE, 2011]. A complete set of base maps have been compiled at 250 m spatial resolution, the gridded data are available at PCRaster format [Karssenberg, et al. 2010] and can be used in models like RUBEM [Mélo Júnior, et al. 2022], SPHY [Terink, et al. 2015], PCR-GLOBWB [Sutanudjaja, et al. 2018], LISFLOOD [Van der Knijff and Younis 2010], WetSpa [Batelaan and De Smedt 2001] and WFLOW

[Köhler, et al. 2006]. This document gives a brief summary of the source of the geo-spatial data and the applied methodology.

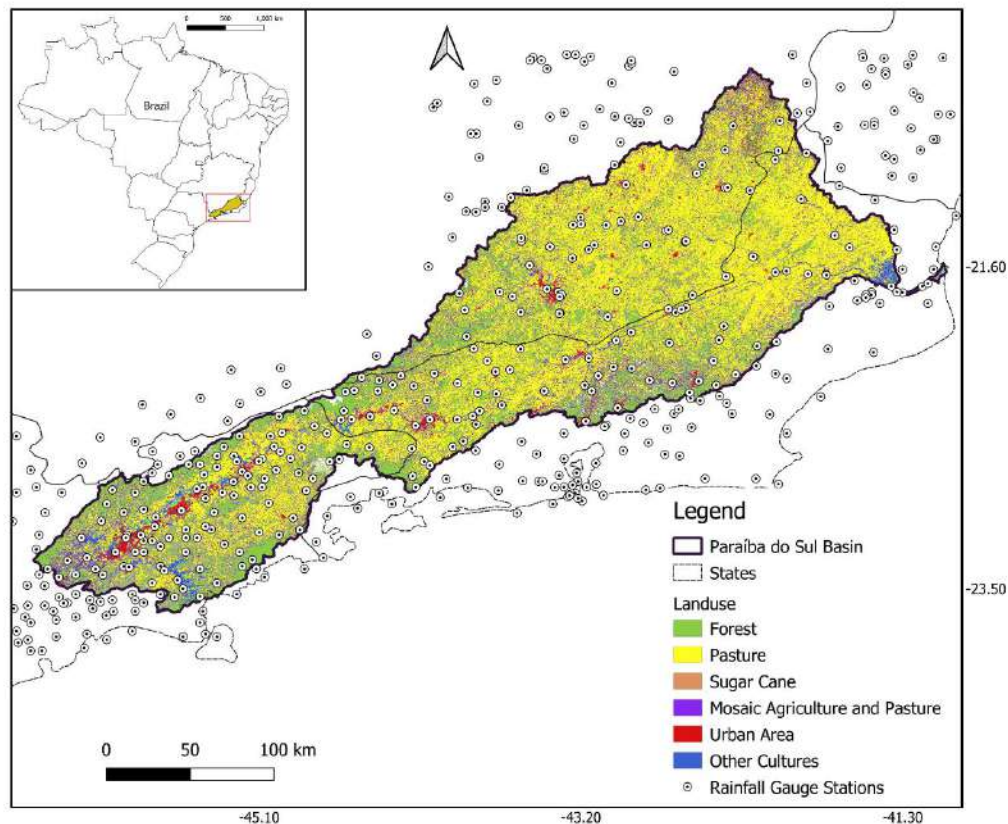


Figure 1. Basin geographic location.

2. Data General Information

- Data Format: GeoTiff, PCRaster raster files and temporal series (*.txt);
- Projection: EPSG:4326 – WGS 84;
- Extent (dd):
 - x_{min} : -46.755;
 - x_{max} : -40.7457;
 - y_{min} : -23.9934;
 - y_{max} : -20.3418.
- Spatial Resolution: 250m;
- Temporal Resolution: monthly;
- Time period: Jan/2000 – Dec/2021;
- Available data:

1. Monthly meteorological raster series of: rainfall [mm/month], potential evapotranspiration [mm/month], evaporation coefficient - k_p [-];
2. Monthly filled images of NDVI [-];
3. Maps of gridded of soil characteristics and tables- wilting point [θ (cm^3/cm^3)], field capacity [θ (cm^3/cm^3)], hydraulic conductivity [mm/month], depth root zone [cm], saturated capacity water content [θ (cm^3/cm^3)], soil bulk density [g/cm^3];
4. Annual land use raster maps and land use parameters tables;
5. Digital elevation model.

3. Meteorological Data

The meteorological data necessary to run the model is summarized in:

- Rainfall;
- Potential Evapotranspiration;
- Class A Pan Coefficient (K_p).

Rainfall was obtained from stations distributed in each basin area, 389 stations were used for the spatialized product (Figure 1), regional vector method was used for filling gaps after a data evaluation of the raw data. Point measures were converted to raster data by kriging interpolation and variogram parameters were optimized using the SciPy-library algorithm in a Python script. For potential evapotranspiration, point values (at station locations) were calculated using the Penman-Montheit method, data from NASA POWER Project were used to fill gaps and kriging interpolation were used to generate the raster series. The Class A pan coefficient was calculated using the wind speed, and relative humidity.

4. Digital Elevation Model

The raster map with elevation has as data resource the global elevation data at 1 arc second spacing NASADEM [NASA 2020]. NASADEM data products were derived from original telemetry data from the Shuttle Radar Topography Mission (SRTM). These NASA project applied multiples methodologies for improve the accuracy of the data, using images from other resources to filling gaps and sinks.

5. Land Use Maps and Parameters

The database available in the Annual Mapping of Land Cover and Land Use Project in Brazil (MapBiomas) - Collection 6 were used as source of data. MapBiomas images are based on machine learning algorithms for classification of land cover, resulting in to 49 classes and subclasses, for Collection 6 products, legend is available at https://mapbiomas-br-site.s3.amazonaws.com/C%C3%B3digos_Clases_Legenda_Cole%C3%A7%C3%A3o_7_-_PT_.docx_1_.pdf. Some parameters associated to each type of land use are available in txt tables, these parameters are related to vegetated, soil bare, open water and

impervious fraction areas and Manning roughness (see Mélo Junior et al. 2022 for more information).

The characterization and analysis of the historical series of watershed coverage (LULC) and the assignment of modeling parameters involved the following steps:

1. Download the series of coverages for the period 2000 – 2021 from Mapbiomas collection 6 (2021 was provide by collection7) through the GEE platform;
2. Preparation of files to PCRaster format;
3. Assignment of features for modeling (area fractions per pixel and Manning roughness).

6. Soil Data

The different soil types in the basin were established from a reclassification of available soil maps from the States of São Paulo, Rio de Janeiro and Minas Gerais. The differentiation of the soils considering their depth and textural class (from semi-detailed surveys) greatly influences their hydrological behavior. The reclassification also sought to simplify the number of soil types, as the great diversity of associations in the surveys made it impossible to calibrate a hydrological model.

The primary data source for obtain the soil parameter was the HYBRAS database [Ottoni, 2018], a hydro-physical database of soils in Brazil, with data on water retention and saturated hydraulic conductivity. Other references to obtain averages for each parameter were also included. Maps and tables are available for the parameter mentioned in section 2. Some type of soils presents in the area are: Argissolos (Ultisols), Cambissolos (Inceptisols), Latossolos (Oxisols), Organossolos (Histosols), Neossolos (Entisols quartzipsamment), Espodossolos (Spodosols) and Gleissolos (Entisols -Aqu-alf-and-ent-ept). Raster maps of each parameter and a table with the legend associated are available in the dataset.

7. NDVI Data

NDVI describes the normalized ratio between the near-infrared and red bands, formally described according to Equation 1:

$$NDVI = \frac{\rho_{nir} - \rho_{red}}{\rho_{nir} + \rho_{red}} \quad 1$$

Where ρ_{nir} corresponds to the value of the near-infrared band and ρ_{red} to the value of the red band of the electromagnetic radiation spectrum. The NDVI index allows monitoring seasonal and interannual changes in vegetation development and activity, being widely used in environmental and natural science research [Jensen, 2009].

Based on the analysis of image availability, the product MODIS - MOD09A1.061 Terra Surface Reflectance 8-Day Global 500 m, was selected to obtain NDVI data. Images collected presented gaps in some areas as a filter for cloud cover and pixel quality were applied.

7.1 Gap-Filling in NDVI images - Proposed Methodology

A simplified methodology for NDVI gap-filling was developed to obtain full monthly images with consistent information linked to annual coverage maps. This association proved coherent. Even if the frequency of coverage is annual, the reading of NDVI monthly reflects the characteristics and state of maturity of the vegetation cover.

The developed methodology consists of 4 steps:

- **Step 1: NDVI image pairing with faults and associated coverage map:** In this step, the monthly NDVI image is associated with the corresponding annual coverage map;
- **Step 2: Obtaining the average NDVI values by coverage class:** From the valid or flawless pixels of the NDVI image, a reading associated with each coverage class of the annual map was taken to obtain the average NDVI values for the month of analysis;
- **Step 3 – Definition of filling criteria:** The criterion for gap-filling in faulty pixels is selected after compiling the monthly average NDVI values for each coverage class. Three criteria were listed for filling gaps, which depend on flawless pixel and coverage availability, in the month under analysis. The criteria are follows:
 - **Criteria 1:** filling pixels with flaws associated with the average NDVI values obtained in the same month, by coverage class;
 - **Criteria 2:** completion with mean NDVI values obtained in the immediately preceding and subsequent months, by coverage class;
 - **Criteria 3:** completion with annual average NDVI values, by coverage class.
- **Step 4 – Filling the gaps:** Finally, monthly NDVI gaps were filled in according to the criterion selected for filling in a group of pixels associated with a given coverage class.

8. Data Availability

The dataset is available at HydroShare web based hydrologic information system <https://www.hydroshare.org/resource/07d089fcc88345b4846f455fa0eced92/>. Additional information about data processing (methodologies, scripts) are available in metadata files in the folders database and at <https://rubem.readthedocs.io/en/latest/preprocessing.html>.

References

- Batelaan, O. and De Smedt, F. (2001) “WetSpaas: A Flexible, GIS Based, Distributed Recharge Methodology for Regional Groundwater Modelling”, IAHS-AISH Publication, p. 11–18.
- Jensen, J. R. (2009) “Sensoriamento Remoto do Ambiente - Uma Perspectiva em Recursos Terrestres”, 2nd ed.; Parêntese: São José dos Campos, BR, ISBN 978-85-60507-06-1.

- Karssenber, D., Schmitz, O., Salamon, P., de Jong, K., and Bierkens, M. F. P. “A software framework for construction of process-based stochastic spatio-temporal models and data assimilation”, *Environmental Modelling & Software* 25(4), p. 489-502, 2010. <https://doi.org/10.1016/j.envsoft.2009.10.004>.
- Kepeng F. and Juncang T. (2021) “Forecasting reference evapotranspiration using data mining and limited climatic data”, *European Journal of Remote Sensing*, 54:2, p. 363-371. <https://doi.org/10.1080/22797254.2020.1801355>.
- Köhler, L., Mulligan, M., Schellekens, J., Schmid, S. and Tobón, C. (2006) “Final Technical Report DFID-FRP Project no. R7991 Hydrological impacts of converting tropical montane cloud forest to pasture, with initial reference to northern Costa Rica”, available at https://assets.publishing.service.gov.uk/media/57a08c3940f0b6497400109a/R7991_FTR.pdf
- Méllo Júnior, A. V., Olivos, L. M. O., Billerbeck, C., Marcellini, S. S., Vichete, W. D., Pasetti, D. M., da Silva, L. M., Soares, G. A. S., Tercini, J. R. B. (2022) “Rainfall Runoff Balance Enhanced Model Applied to Tropical Hydrology”. *Water*, 14, 1958. <https://doi.org/10.3390/w14121958>.
- NASA JPL (2020) “NASADEM Merged DEM Global 1 arc second V001”. NASA EOSDIS Land Processes DAAC, https://doi.org/10.5067/MEaSURES/NASADEM/NASADEM_HGT.001.
- OECD (2017). “The Paraíba do Sul River basin, in Water Charges in Brazil: The ways Forward”, OECD Publishing, Paris. <https://doi.org/10.1787/9789264285712-10-en>.
- Otoni, M. V. (2018) “HYBRAS Hydrophysical Database for Brazilian Soils: Banco de Dados Hidrofísicos em Solos no Brasil para o Desenvolvimento de Funções de Pedotransferências de Propriedades Hidráulicas”: Ver. 1.0: Relatório; CPRM, Ed.; Rio de Janeiro, BR, ISBN 978-85-7499-370-6.
- Sutanudjaja, E. H., Van Beek, R., Wanders, N., Wada, Y., Bosmans, J., Drost, N., Van der Ent, R., De Graaf, I., Hoch, J., De Jong, K., Karssenber, D., López López, P., Peßenteiner, S., Schmitz, O., Straatsma, M., Vannamettee, E., Wisser, D., and Bierkens, M. (2018) “PCR-GLOBWB 2: a 5 arcmin global hydrological and water resources model”, *Geosci. Model Dev.*, 11, p. 2429–2453.
- Terink, W., Lutz, A. F., Simons, G. W. H., Immerzeel, W. W., Droogers, P. (2015) “SPHY v2.0: Spatial Processes in Hydrology”, *Geosci. Model Dev.*, 8, p. 2009–2034. <https://doi.org/10.5194/gmd-8-2009-2015>.
- Van Der Knijff, J. M., Younis, J. and De Roo, A. P. J. (2010) “LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation”, *International Journal of Geographical Information Science*, 24(2), p. 189-212. <https://doi.org/10.1080/13658810802549154>.
- Xavier, A. C., King, C. W. and Scanlon, B. R. (2016) “Daily gridded meteorological variables in Brazil (1980–2013)”. *International Journal Climatology.*, 36, p. 2644-2659. <https://doi.org/10.1002/joc.4518>.

Mobility Database of the Metropolitan Region of Belo Horizonte, Brazil

Lívia Rodrigues Tomás¹, Leonardo B. L. Santos¹

¹ Brazilian National Center for Monitoring and Early Warning of Natural Disasters (Cemaden) – Estrada Dr. Altino Bondensan, 500 - 12247-016 - São José dos Campos – SP - Brazil

liviatomas@gmail.com, santoslbl@gmail.com

Abstract. *This paper describes the mobility database of the metropolitan region of Belo Horizonte, Brazil. It comprises the official data from the origin and destination survey carried out in 2012, 2019 and 2021 and it is available on GitHub online repository. It contains vector files (shapefiles) and alphanumeric files (.txt and .xlsx). Shapefiles refers to: municipalities, macro mobility unit, homogeneous areas, traffic zones and census sectors' boundaries and the road system. The geography coordinate system is EPSG 4326 – SIRGAS2000.*

Metropolitan Region of Belo Horizonte (MRBH)

Located in the central region of the state of Minas Gerais, with the capital of Minas Gerais as its hub, the Metropolitan Region of Belo Horizonte (MRBH) is currently composed of thirty-four municipalities: (1) Baldim, (2) Belo Horizonte, (3) Betim, (4) Brumadinho, (5) Caeté, (6) Capim Branco, (7) Confins, (8) Contagem, (9) Esmeraldas, (10) Florestal, (11) Ibirité, (12) Igarapé, (13) Itaguara, (14) Itatiaiuçu, (15) Jaboticatubas, (16) Juatuba, (17) Lagoa Santa, (18) Mário Campos, (19) Mateus Leme, (20) Matozinhos, (21) Nova Lima, (22) Nova União, (23) Pedro Leopoldo, (24) Raposos, (25) Ribeirão das Neves, (26) Rio Acima, (27) Rio Manso, (28) Sabará, (29) Santa Luzia, (30) São Joaquim de Bicas, (31) São José da Lapa, (32) Sarzedo, (33) Taquaraçu de Minas e (34) Vespasiano, as shows Figure 1.

According to data from the 2010 Census, the MRBH has almost five million inhabitants, of which 98% are urban. The population density is around 500 inhab./Km². With a GDP of BRL 120,833,976 and a GDP per capita of BRL 828,821.73, the RMBH is the third largest metropolitan region in Brazil, in economic and population terms (MRBH Agency, 2012).

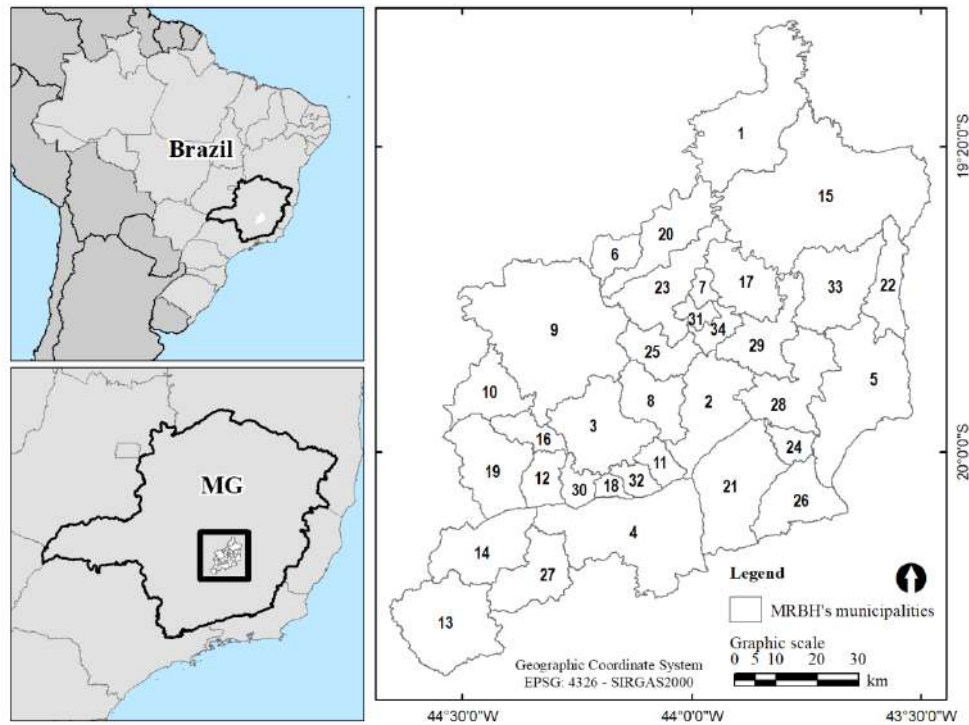


Figure 1. Location of Metropolitan Region of Belo Horizonte – MRBH

Origin and Destination Survey

The Origin and Destination (OD) Survey is the main data source for knowing the dynamics of people's movements in the researched area, associating the characteristics of the trips with socioeconomic variables.

In 2012, the Belo Horizonte Metropolitan Region Development Agency carried out a household OD Survey in all 34 municipalities in the MRBH to identify the trips patterns of its inhabitants. The official data of this OD Survey can be found at RMBH Agency site (RMBH Agency, 2022).

The spatial unit adopted for carrying out the research was the homogeneous area, defined as a set of aggregated census sectors, according to socioeconomic and occupation criteria. A total of 1,289 Homogeneous Areas (HA) were defined, 554 of which in the municipality of Belo Horizonte and 735 distributed in the other municipalities of the MRBH. Each HA aggregates, on average, three census sectors. This unit is geared towards addressing urban issues.

A spatial unit called Macro Mobility Unit - MMU was also created to analyze specific characteristics of urban mobility. 67 MMUs were defined, resulting from the aggregation of HAs. Figure 2 shows their boundaries.

In 2019/2021, they collected information from the mobile phone network to develop an updated OD matrix for the MRBH (RMBH Agency, 2022). This matrix expresses the daily number of trips among all OD pairs, classifying them as to their reason, period of travel, and travelers' sociodemographic profile.

Two periods of 20 days were considered for the analysis of trips, obtaining the average volume of trips per business day and holidays. The first period was from November 01 to 20, 2019. The second period was from May 01 to 20, 2021.

The Traffic Zone (TZ) was the spatial unit used to report the results of the OD matrix. It is defined as a set of aggregated HAs whose land use and sociodemographic profile are similar, resulting in 393 polygons. There is a relational table between TZ and HA. Figure 3 shows the TZs boundaries.

Database

This database (DB) is available on GitHub online repository (Tomás, 2022). It contains vector files (shapefiles) and text files (.txt). Shapefiles refers to: municipalities, MMUs (2012), HAs (2012), TZs (2019/2021) and census sectors' boundaries (2010) and the road system. The geography coordinate system is EPSG 4326 – SIRGAS2000.

Text files consider the number of trips between OD pairs. Being 'trip' the movement of one person, for any specific reason, with the use of one or more means of transport, and may be composed of one or more transfers. Therefore, each trip has an origin, a destination, a reason and a mode of transport. In this DB, the origin and destination are one of the MMUs or HAs (for 2012) and TZs for 2019/2021. The 2019 and 2021 matrices consider trips carried out on business days for comparison with the OD 2012.

Matrices consider metropolitan trips, that is, trips starting and ending in the MRBH, which in turn are divided into internal trips (intrazone trip), trips for which origin and destination are both located in the same zone; and external trips (interzone trip), which are trips between different zones. The DB also presents the produced trips, that is, the trips that originated in the same zone with different destinations; attracted trips, that is, trips with different origins and destinations in the same zone; as well as total trips.

Considering the total trips in 2012, Figure 2A shows a choropleth map by MMU. It is possible to visualize the relationship of territorial proximity of the MMU with the highest volume of trips, Center-South, with the others. In order to compare the pattern of trips between the 2012 and 2019/2021 surveys, the number of trips, originally aggregated by HA, was also aggregated by TZ, as Figure 2B shows. Figure 3 shows the total trips by TZ in (a) 2019 and (b) 2021. The class ranges are the same for both years; however, the features values of each year were used at the extremes (minimum and maximum values).

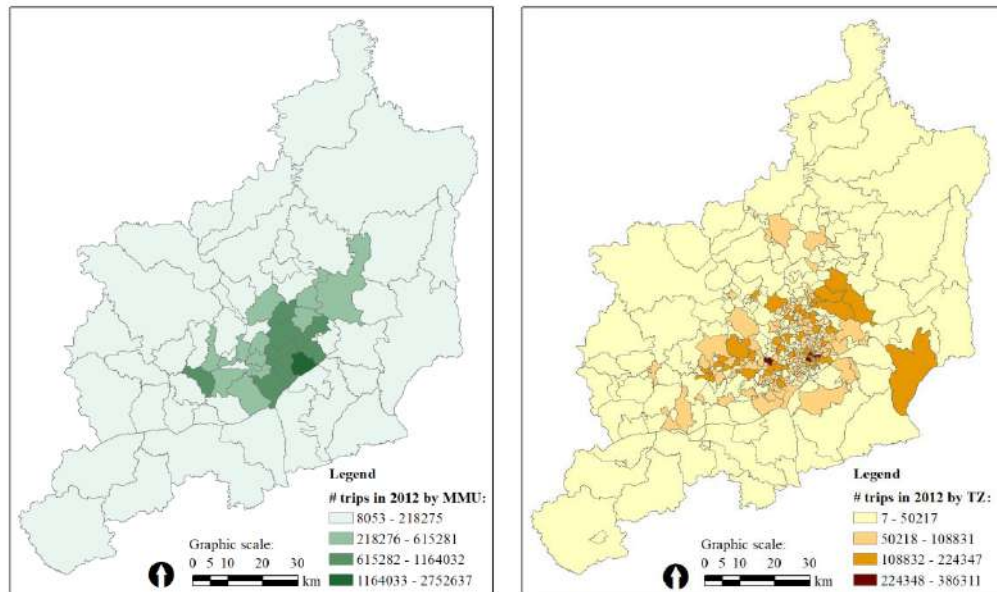


Figure 2. Number of trips in 2012 by (a) MMU and (b) TZ.

Source: Authors' elaboration, based on data from the 2012 OD Survey.

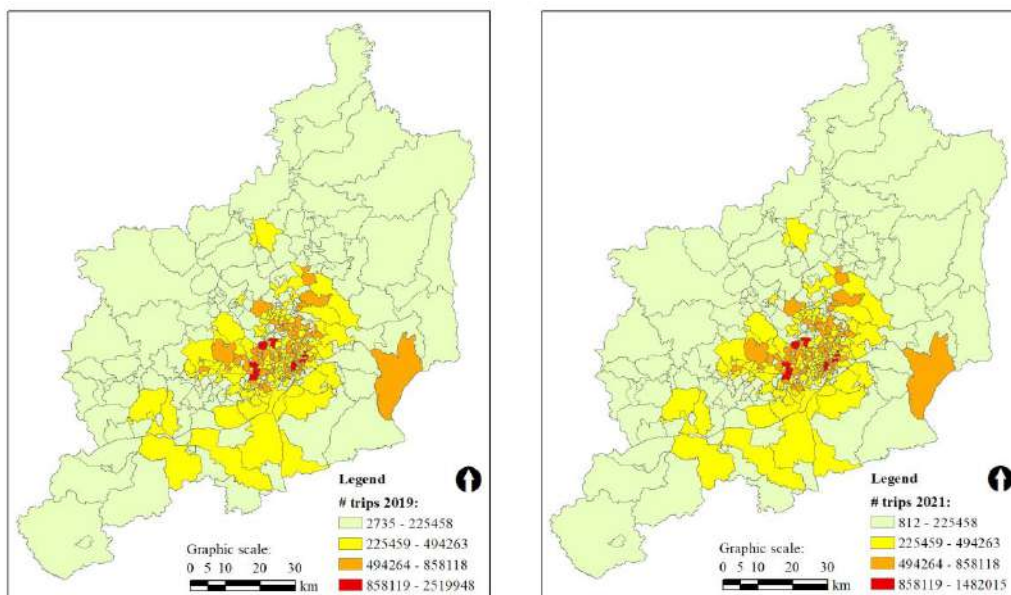


Figure 3. Number of trips by TZ in (a) 2019 and (b) 2021.

Source: Authors' elaboration, based on data from the 2019/2021 OD Survey.

In 2012 there were a total of 23,137,457 trips in MRBH, and TZs with the highest number of trips were, respectively, 126, 69, 233, 11, and 109. In 2019 there were a total of 162,869,637 trips, and TZs with the highest number of trips were 109, 126, 69, 35, and 236, respectively. In 2021 there were a total of 122,424,794 trips and

TZs with highest number of trips were 233, 109, 126, 29, and 238, respectively. These TZs belongs to Belo Horizonte and Contagem municipalities.

Final remarks

Although it is older, the 2012 OD research is home-based and it has a greater number of variables. The 2019/2021 surveys use mobile phone networks, which makes the research faster and cheaper, but it has fewer variables and some of them are estimated. The combined use of these two types of research allows us to understand changes in the pattern of trips over time. Observing the total trips in MRBH by year, it is possible to note that there was almost a 604% growth between 2012 and 2019; and there was almost a 25% decrease from 2019 to 2021. When we look at the highest number of trips by TZ per year (i.e. in 2012 the number of trips for TZ 126; in 2019, the number of trips for TZ 109; and in 2021 the number of trips for TZ 233), that there was a 15% growth in the number of trips between the years 2012 and 2019; and a 41% decrease from 2019 to 2021. This decrease is explained by the change in the pattern of travel caused by the SARS-CoV-1 pandemic, in which many people started working from home.

References

- RMBH Agency. (2022) “Origin and Destination Survey”, <http://www.agenciarmbh.mg.gov.br/pesquisa-od/>.
- Tomás, L. R. (2022) “Mobility database of Metropolitan Region of Belo Horizonte, Brazil”, <https://github.com/liviatomas/RMBH>.

An urban flood database for Disaster Risk Reduction studies in São Paulo, Brazil

Lívia Rodrigues Tomás¹, Leonardo B. L. Santos¹

¹ National Center for Monitoring and Early Warning of Natural Disasters (Cemaden) – Estrada Dr. Altino Bondensan, 500 - 12247-016 - São José dos Campos – SP - Brazil

liviatomas@gmail.com, santoslbl@gmail.com

Abstract. *This paper describes the urban flood database of an area in the São Paulo municipality, Brazil. The database is organized into input, processed, output and auxiliary data, and it contains vector files (shapefiles) and alphanumeric files. The input data have their original spatial topology. The processed data presents the spatialization of alphanumeric data as well as generated by-products. The output data presents the variables aggregated in a regular grid. It is available on GitHub online repository. The geography coordinate system is EPSG 4326 – SIRGAS2000.*

Background and summary

Flooding is one of the most frequent and damaging natural disaster affecting millions of people around the globe (Jongman et al., 2015a). It can be defined as hydrological phenomena caused by an excess capacity of surface runoff and urban drainage systems forming accumulations of water in impermeable areas (CENAD, 2014). In Brazil, the hydrometeorological disasters accounted for 84% of the unsheltered from 2000 and 2019 (Perez et al., 2020).

For flood detection, data collected by gauge systems in rivers, remote sensing (Revilla-Romero et al., 2014), on-site surveys or other sources, such as media reports and government agencies (Jongman et al., 2015b), are normally used (de Bruijn et al., 2019).

The Flood Alert System of São Paulo (SAISP) generates reports at two-hour intervals for each rain event, informing about its progress, power, and duration. The municipality also counts on the Climate Emergency Management Center (CGE) to monitor flood areas in the city of São Paulo and mitigate the effects of the occurrences. CGE uses radar, rain gauge measurements and flood reports provided by the Municipal Civil Defense Coordination (COMDEC) and the Traffic Engineering Company (CET) to monitor and to emit flood alerts (López and Rodríguez, 2020).

The CGE keeps the flood point records on its website (CGE, 2022) and anyone can access the data. It is necessary to search by date to access the history of floods. The search result is shown on the screen and it contains information on the location of the flood, the start and end time of the road interdiction and a classification of the point (whether it is active or inactive and whether the road is passable or impassable). It is only possible to consult one date at a time and the system does not export the result.

The occurrence of flooding with the prohibition of an important road has a great impact on people's lives, and may hinder or preclude to return to residence. In this

sense, accurate and timely information on the geographic location and impacts of ongoing floods is essential for disaster response (Jongman et al., 2015b).

The potential of a hazard to become a disaster depends on the degree of exposure of the population, vulnerability, probability of occurrence, amongst others (Dickson et al., 2012). Thus, these variables must be included in the disaster risk reduction management.

Study area and datasets

The study area (Figure 1) belongs to Tamanduateí river basin in São Paulo municipality. It is a built-up area with impervious surfaces with 115.3 km², and 1,637,265 inhabitants (IBGE, 2016). It is an area with great attraction for trips, where 4,760,743 people work or study there (METRO, 2019).

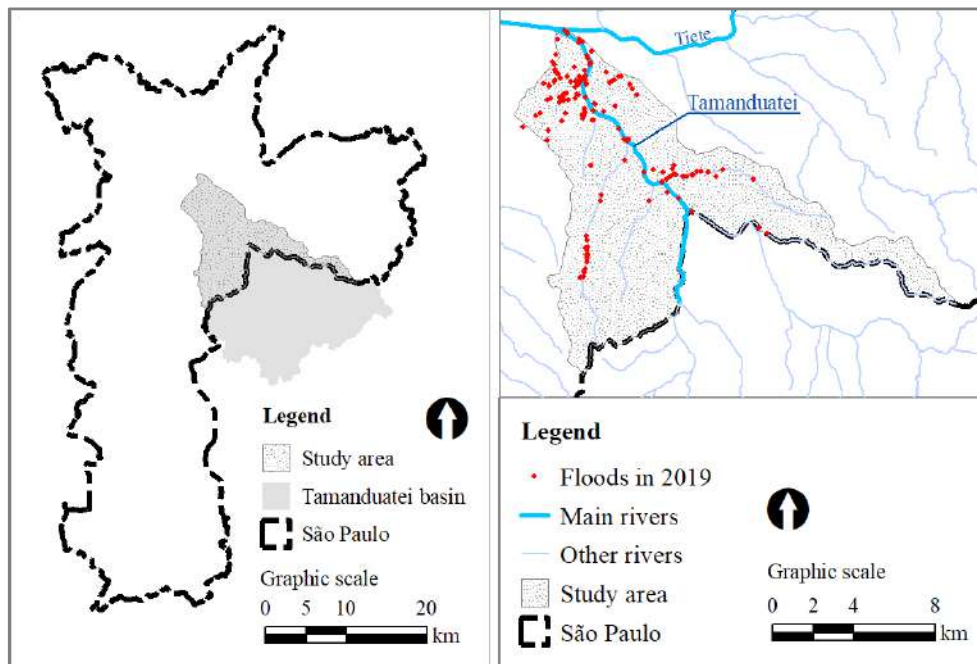


Figure 1. Study area location

To estimate the risk of flooding, and the variables associated with it, we created a geographic database considering socioeconomic data, occurrence of flooding, urban mobility, soil relief and characteristics of the road system. These input data have different sources and spatial topologies, and they are available as follows:

- (a) Regular grid from 2010 Brazilian census (Fundação Instituto Brasileiro de Geografia e Estatística - IBGE, 2016);
- (b) Inhabitants by cell grid (Fundação Instituto Brasileiro de Geografia e Estatística - IBGE, 2016);
- (c) Geographic coordinates of workplaces and the number of people who work at each location from the Origin Destination Research Database (OD Database) (METRO, Companhia do Metropolitano de São Paulo, 2019);

- (d) Geographic coordinates of educational institutions and the number of people who study at each location from the OD Database (METRO, Companhia do Metropolitano de São Paulo, 2019);
- (e) Flood registries in 2019 obtained from the Emergency Management Center of the city of São Paulo - CGE website (Climate Emergency Management Center - CGE, 2021);
- (f) Road system vector data from the Center for Metropolitan Studies - CEM (2021);
- (g) SRTM Digital Elevation Model - DEM (Farr et al., 2007).

We generate new layers, after processing the input data. The alphanumeric data of items (c), (d), and (e) were used to generate vector files (shapefiles) with points representing the location of workplaces, education institutions and floods occurrence respectively. The SRTM DEM model (item g) was used to obtain the Height Above the Nearest Drainage (HAND) raster. Roads vector was used to calculate the roads vulnerability index.

The processed data was then aggregated in the IBGE regular grid considering their quantitative value. Each grid cell received these attributes, considering their spatial location and these approaches:

- Sum of commuters and resident population, resulting in the exposed population;
- Sum of road closure time, resulting in local vulnerability;
- Maximum value of the node vulnerability of the intersected lines, resulting in network vulnerability;
- Minimum HAND value among those that intersect it.

These variables aggregated in the grid cell were then classified and combined in three classes: moderate, high and very high, resulting in the flood risk map and its related components. The detailed methodology and all steps are described in Tomás et al., 2022. The output is the regular grid with five attributes: exposure, vulnerability, flood susceptibility, potential impact and flood risk.

The database is available at *Github* repository (Tomás, 2022), organized into the following folders: input, processed, output, and auxiliary.

Final Remarks

For most cities in developing countries, collecting reliable and timely data is a challenging task, especially from a multidisciplinary point of view. Integrating knowledge from multiple fields better reflects the geographic environment in which various natural processes, human activities, and information interactions exist. In this sense, methodologies comprised of hydrology, mobility, and Geographic Information Systems (GIS) have proved effective in many case studies to model complex urban systems. Furthermore, GIS are of great help to analyze data and produce meaningful information for urban management and resilience planning.

This database combines elements from natural geographies, such as hydrological indexes, and elements from human geographies, such as urban mobility data, in both

cases using different sources and geographic units. It was originally created to map the risk of flooding in the city of São Paulo, but allows its use for other areas of knowledge, since it has variables of different natures. We hope that it can contribute to sustainable urban planning and can be added to other perspectives on the city.

References

- Center for Metropolitan Studies - CEM. (2021). “Digital Geo-referenced Cartographic Base of Roads System in the Metropolitan Region of São Paulo – RMSP - 2020 Edition”. <https://centrodametropole.fflch.usp.br/pt-br/download-de-dados>
- CGE - Climate Emergency Management Center. (2022). “Flood records in São Paulo, Brazil”. <https://www.cgesp.org>
- de Bruijn, J. A., de Moel, H., Jongman, B., de Ruiter, M. C., Wagemaker, J., & Aerts, J. C. J. H. (2019). A global database of historic and real-time flood events based on social media. In *Sci Data*, 6, 311.
- Dickson, E., Baker, J. L., Hoornweg, D., & Tiwari, A. (2012). Urban risk assessments: An approach for understanding disaster
- Farr, T. G., et al. (2007). The shuttle radar topography mission. In *Reviews of Geophysics*, 45(2), RG2004. https://doi.org/10.1029/2005R_G000183
- Fundação Instituto Brasileiro de Geografia e Estatística - IBGE. (2016). “Grade Estatística Censo 2010”. <https://www.ibge.gov.br/geociencias/downloads-geociencias.html>
- Jongman, B. et al. (2015a). Declining vulnerability to river floods and the global benefits of adaptation. In *Proc. Natl. Acad. Sci.*, 112, E2271-E2280.
- Jongman, B., Wagemaker, J., Revilla-Romero, B., & de Perez, E. C. (2015b). Early Flood Detection for Rapid Humanitarian Response: Harnessing Near Real-Time Satellite and Twitter Signals. In *ISPRS Int. J. Geo-Information*, 4, 2246–2266.
- López, A. S. V. and Rodriguez, C. A. M. (2020). Flash Flood Forecasting in São Paulo Using a Binary Logistic Regression Model. In *Atmosphere*, 11, 473.
- METRO, Companhia do Metropolitano de São Paulo. (2019). “Pesquisa Origem-Destino 2017”. <http://www.metro.sp.gov.br/pesquisa-od/>
- Ministério de Integração Nacional, Secretaria Nacional de Defesa Civil & Centro Nacional de Gerenciamento de Riscos e Desastres – CENAD. (2014). “Anuário Brasileiro de Desastres Naturais”.
- Perez, L. P., Rodrigues-Filho, S., Marengo, J. A., Santos, D. V., Mikosz, L. (2020). Climate change and disasters: analysis of the Brazilian regional inequality. In *Sustainability in Debate*, 11 (3), 260-277
- Revilla-Romero, B., Thielen, J., Salamon, P., De Groeve, T., & Brakenridge, G. R. (2014). Evaluation of the satellite-based global flood Detection System for measuring river discharge: Influence of local factors. In *Hydrol. Earth Syst. Sci.*, 18, 4467-4484.

Tomás, L. R. (2022) “Urban flood database of an area of São Paulo municipality”,
https://github.com/liviatomas/floods_saopaulo.

Tomás, L. R., Soares, G. G., Jorge, A. A. S., Mendes, J. F., Freitas, V. L. S. & Santos,
L. B. L. (2022a). Flood risk map from hydrological and mobility data: A case study
in São Paulo (Brazil). In *Transactions in GIS*, 26, 2341– 2365.
<https://doi.org/10.1111/tgis.12962>

Towards generating a traffic slowness geospatial dataset of São Paulo city

Jeferson F. Mendes¹, Rogério G. Negri¹, Leonardo B. L. Santos²

¹Department of Environmental Engineering, Sciences and Technology Institute, São Paulo State University (UNESP), São José dos Campos, São Paulo, Brazil.

²Center for Monitoring and Early Warning of Natural Disasters (CEMADEN), São José dos Campos, São Paulo, Brazil.

{jeferson.feitosa8,santoslbl}@gmail.com, rogerio.negri@unesp.br

Abstract. *The city of São Paulo is known for its extensive fleet of vehicles, which, combined with extreme precipitation and flooding events, makes traffic increasingly chaotic. The present work comprises an extracting, transforming and loading (ETL) process to generate a database of 2019 traffic slowness in São Paulo integrated with variables that may influence traffic, such as rainfall, floods, accidents and socioeconomic features. This dataset is expected to support more profound analysis and help in better traffic planning and decision-making.*

1. Introduction

It is well known that São Paulo faces a serious mobility problem in its territory. With more than 20 million inhabitants, the municipality is one of the largest and most congested metropolitan regions in the world. More than 30 million motorized trips are made on average in São Paulo, with an average length of 8 km and duration of 51 min [IBGE 2013].

One of the main causes of this issues is the fact that the number of vehicles has been increasing significantly in recent years. According to São Paulo State Traffic Department (CET, 2021), the São Paulo fleet in January 2021 was around 8.9 million cars, which is equivalent to 8.3% of the Brazilian fleet. In the last 9 years, the car/inhabitant ratio has increased by 71%. However, the level of congestion in the transport network not only varies due to the recurring congestion at peak times due to the number of vehicles, but also due to the non-recurring congestion caused by unexpected events, such as accidents and disasters [Chikaraishi et al. 2020].

Over the last decades, social-environmental disasters have been increasing both in frequency and intensity, probably due to a climate change scenario [World Meteorological Organization et al. 2014]. According to the United Nations International Strategy for Disaster Reduction (2017), between 1998 and 2017, climate disasters were responsible for 91% of all 7,255 disasters recorded worldwide. In São Paulo, capital city, the situation is not different. Despite the uncertainties that permeate long-term climate projections, Recent studies already point to an increase in the frequency and intensity of rainfall in the capital of the Mega city [Santos and Haddad 2014].

According to Yi (2021), for most events, a two-hour rainfall can impact congestion over a period of up to five hours. Hence, the current São Paulo's chaotic traffic tends to become even more problematic in flood events [Lu et al. 2020]. However, besides

meteorological events, there are many other variables that may possibly impact the traffic slowness, including structural, environmental and socioeconomic features.

Therefore, face with the dimension that the problem has reached and the expected aggravation of this situation, it is necessary that the public managers adopt more effective postures in facing the problem, since this situation compromises the economy and the population well-being [Cintra 2014],[Thomas et al. 2018]. An important step to deal with this problem comprises the development of spatio-temporal traffic databases. This information is essential to build models to predict the traffic slowness as function of local variables, and consequently support better urban-planning and traffic management [Van Arem et al. 1997].

In this work presents a geospatial database building process for São Paulo city regarding the traffic slowness in 2019. Besides traffic slowness measurements, the database includes variables that may possibly impact traffic, such as quantity and duration of floods, precipitation, number of car accidents and the socioeconomic local profile.

2. Material and Methods

2.1. Study area, data and tools

The study area comprises a portion of São Paulo city (46°40'41"W 23°34'41"S), state of São Paulo, in the Southeastern Brazil. São Paulo is the biggest city in the "São Paulo Metropolitan Region" (SPMR) with about 1521 km² of area and approximately 12.1 million inhabitants. The SPMR comprises 39 cities and stands as the largest industrial complex and urban concentration in Latin America [Barros and Conde 2017].

The traffic data is measured and made available by traffic engineering company (CET). Registers are made every 30 and comprises the main city corridors in the downtown area and the main circulation axes. Currently, 800 km of roads are monitored, which corresponds to about 5% of the total highway network.

Rain data was obtained through the "São Roque weather Radar", a Doppler S-Band Radar (2.7–2.9 GHz) located in the city of São Roque at coordinates 47°05'39"W and 23°36'07"S and altitude of approximately 1147 m. Data is available in reflectivity (dbz) units, which were converted to precipitation (mm/h) and aggregated to represent hourly accumulated rain. Flood data source is the Emergency Management Center (CGE), an organization for monitoring weather conditions and reports *where*, *when* and *how long* a flood lasted. Other variables include the traffic accidents, obtained from São Paulo state government, and socioeconomic feature projection for 2020 from Michael Bouer Institute.

The data exploration, visualization and modeling were carried out using the Python programming language. Furthermore, it was also employed the Pandas, Numpy, Scipy, Matplotlib, Scikit-Learn and Tensor Flow data science libraries. Additionally, the ArcGIS Pro software with ArcPy library was used to manipulate spatial data.

2.2. Pre-processing (geocoding)

One of the most important methodological step in this study was geocoding (i.e., georeferencing) the roads where traffic is measured, so that spatial analysis can be applied and location-based features can be considered on the database construction. This is due to

the fact that the source slowness data is in non spatial tabular format. Each register in the database specifies the measured slowness, time of day, corridor name and road stretch description. However, both corridor name and stretch are pattern-less text fields, which makes nearly unfeasible to use approaches such as geocoding [Rashidian et al. 2017].

The approach used to geolocate the data consisted primarily in aggregating measurements that were made at the same time and in the same corridor, then summarizing it's total slowness. This means that for each time and corridor pair, the slowness of all stretches are summed up. This allowed integrating the slowness dataset with a georeferenced São Paulo roads shapefile made available by the city government.

The merge of slowness and roads data was made using corridor name as the matching field. One record in the slowness data may correspond to many records in the roads shapefile, since the slowness data uses a more general nomenclature for the corridor. For example, one of the corridors is “Radial Leste”, an informal name for a stretch composed of many avenues. On the roads shapefile, “Radial Leste” record doesn’t exist, however the individual avenues that compose it does. Some of this cases were matched manually, with assist of the Levenshtein Distance similarity algorithm [Haldar and Mukhopadhyay 2011], to allow even partially different registers (regarding to the corridor name) to be matched.

2.3. Data processing

Once all 2019 data was gathered, it was integrated along with traffic slowness georeferenced data, using spatial influence and temporal-context filters. Spatial influence refers to the minimum distance between the data point (i.e flood, rain or accident) and a road. The temporal-context filter select data points according to the minimum time lapse (difference) between the occurrence and the instant when the slowness was measured.

Table 1. Spatial and temporal parameters considered to merge data

Data type	Spatial influence	Temporal-context
Rain	1000 m	1 hour before
Floods	500 m	During
Accidents	300 m	1 hour before
Socioeconomic	500 m	Not used

Rain data was aggregated such that each record contains the average precipitation (mm) for a specific corridor and date. The same logic was applied to floods and accidents, but those contain the number of events for each corridor and date. Finally for each corridor, socioeconomic data contains the population density, household totals, average household size, primary education index, secondary education index, superior education index and purchasing power.

3. Results

As result, it was generated a 167000 records vector file, in *shapefile* format (Figure 1), where each register contains the corridor, date and measured traffic slowness. This dataset was merged with Rain, floods, accidents and socioeconomic data using corridor and date as primary keys, resulting in 4 additional spatial datasets (Figure 2).

It is worth mentioning, as a main error source, possible mismatches between road names during the initial merging process between slowness data and roads vectors. This occurs mainly by the subjectivity of some corridor names, which made it difficult to determine which roads are associated with each corridor. All of the data and code used to generate it are publicly available on GitHub [Mendes 2022]

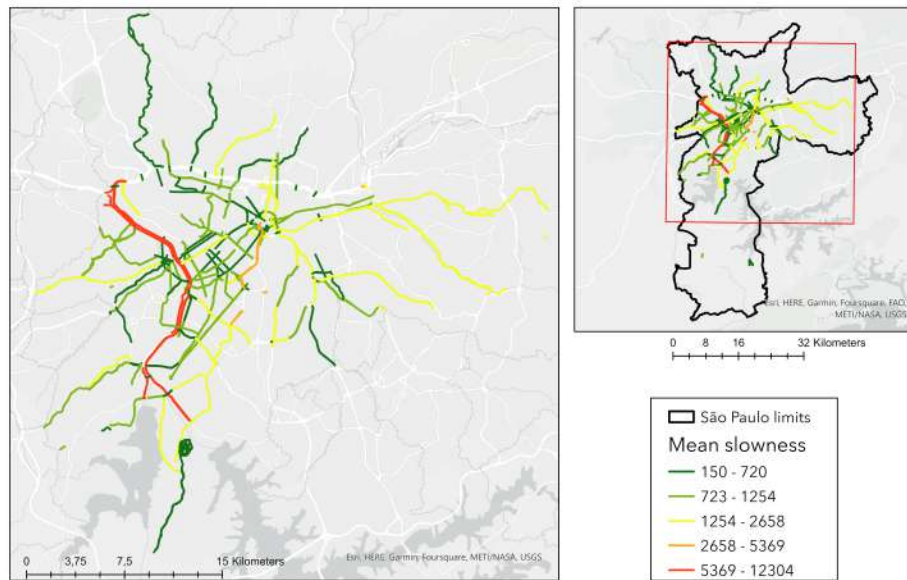


Figure 1. Average slowness in the corridors monitored by CET.

4. Conclusion

Traffic slowness in the city of São Paulo has been increasingly growing due to the high number of vehicles and increase in floods frequency and intensity. The traffic monitoring, made by Traffic Engineering Company(CET) covers the main city corridors, however it is not a spatial data. This makes difficult to perform deeper analysis and integrate slowness with other sources of data, in order to improve planning and decision making. Motivated by this issue, this work created a geospatial database of traffic slowness in the city of São Paulo. Features regarding the measured precipitation, number of floods, car accidents and socioeconomic variables were included in this database.

With the resulting dataset it was possible to highlight features that influence and helps explaining traffic slowness. Future perspectives includes further exploring traffic data and related variables, in order to understand it's spatial and temporal patterns and possibly generate a model to predict traffic slowness in São Paulo city.

References

Barros, M. T. L. and Conde, F. (2017). *Urban Flood Warning System Social Benefits*, pages 642–653.

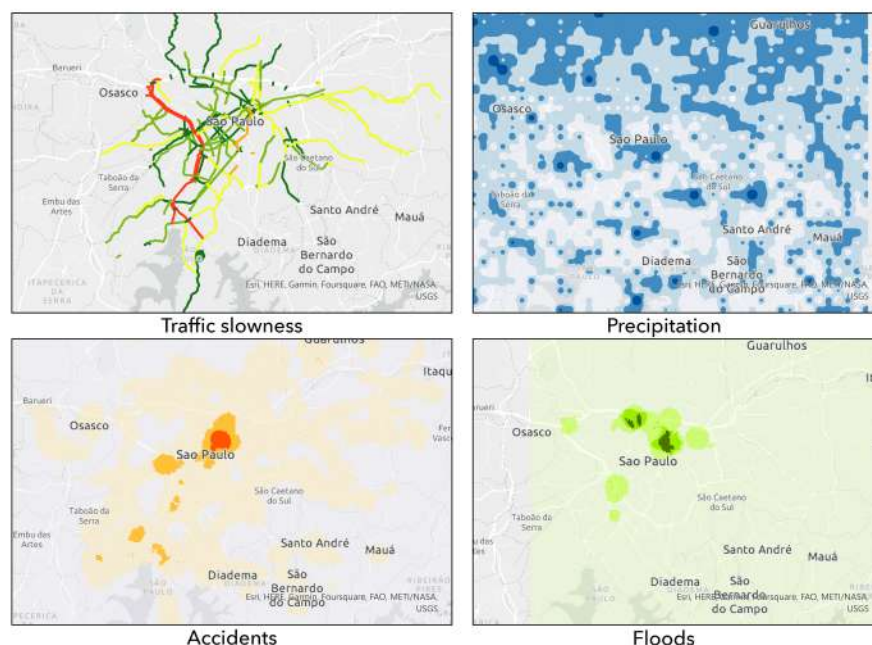


Figure 2. Additional datasets (likely related to traffic) included in the database

- Chikaraishi, M., Garg, P., Varghese, V., Yoshizoe, K., Urata, J., Shiomi, Y., and Watanabe, R. (2020). On the possibility of short-term traffic prediction during disaster with machine learning approaches: An exploratory analysis. *Transport Policy*, 98:91–104.
- Cintra, M. (2014). Os custos do congestionamentos na cidade de são paulo. <https://bibliotecadigital.fgv.br>.
- Haldar, R. and Mukhopadhyay, D. (2011). Levenshtein distance technique in dictionary lookup methods: An improved approach. *Computing Research Repository - CORR*.
- IBGE (2013). Pesquisa nacional por amostra de domicílios. <https://biblioteca.ibge.gov.br/visualizacao/livros/liv94414.pdf>.
- Lu, H., Chen, M., and Kuang, W. (2020). The impacts of abnormal weather and natural disasters on transport and strategies for enhancing ability for disaster prevention and mitigation. *Transport Policy*, 98:2–9.
- Mendes, J. (2022). Traffic slowness prediction in são paulo city. <https://github.com/jeference/traffic-slowness-prediction.git>.
- Rashidian, S., Dong, X., Avadhani, A., Poddar, P., and Wang, F. (2017). Effective Scalable and Integrative Geocoding for Massive Address Datasets. In *Proceedings of the 25th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 1–10, Redondo Beach CA USA. ACM.
- Santos, E. T. d. and Haddad, E. A. (2014). Mapeamento das perdas econômicas potenciais dos pontos de alagamento do município de São Paulo, 2008-2012. *Ambiente & Sociedade*, 17(4):77–94.

- Thomas, T., Mondschein, A., Osman, T., and Taylor, B. D. (2018). Not so fast? Examining neighborhood-level effects of traffic congestion on job access. *Transportation Research Part A: Policy and Practice*, 113:529–541.
- Van Arem, B., Kirby, H. R., Van Der Vlist, M. J., and Whittaker, J. C. (1997). Recent advances and applications in the field of short-term traffic forecasting. *International Journal of Forecasting*, 13(1):1–12.
- World Meteorological Organization, Etienne, C., Williams, M., Castonguay, S., Golaraghi, M., Guha-Sapir, D., Below, R., Jarraud, M., and Centre for Research on the Epidemiology of Disasters (2014). *Atlas of mortality and economic losses from weather, climate, and water extremes (1970-2012)*. OCLC: 884992156.

Q-CleanCAR: Um complemento QGIS para remoção guiada de sobreposições de polígonos de imóveis rurais do Cadastro Ambiental Rural (CAR)

**Maíra Ramalho Matias¹, Gisele Milare^{1,2}, Maria Isabel Sobral Escada¹,
Antônio Miguel Vieira Monteiro¹**

¹Laboratório de Investigações em Sistemas Socioambientais (LISS)
Instituto Nacional de Pesquisas Espaciais (INPE)
Caixa Postal 515 12.227-010
São José dos Campos SP Brasil

²Superintendência Regional do Incra no Estado do Tocantins
Instituto Nacional de Colonização e Reforma Agrária (INCRA)
77.006-336 Palmas TO Brasil

{maira.matias,gisele.milare,isabel.escada,miguel.monteiro}@inpe.br

Abstract. *The self-declared database of the Rural Environmental Cadastre (CAR) has been used for several scientific studies and environment policies. However, these data have inconsistencies, such as geometric overlap among registered properties, which hinders its use in Geographic Information Systems (GIS). This paper presents the "Q-CleanCAR" plugin for the QGIS software, a tool that apply a semi-automatic method for cleaning, classifying of properties by size, removing and registering of overlapping polygons of the properties declared in the CAR.*

Resumo. *A base de dados do Cadastro Ambiental Rural (CAR), de registro autodeclarado, tem sido utilizada para diversos estudos científicos e em políticas ambientais. Entretanto, esses dados apresentam inconsistências, como a sobreposição geométrica entre os imóveis cadastrados, o que dificulta sua utilização em Sistemas de Informação Geográfica (SIGs). O presente artigo apresenta o complemento "Q-CleanCAR" para o software QGIS, uma ferramenta que aplica um método semi-automático de limpeza, classificação de propriedades por tamanho, remoção e registro de sobreposição dos polígonos dos imóveis declarados no CAR.*

1. Introdução

O CAR é um registro público eletrônico autodeclarado de âmbito nacional, obrigatório a todas as propriedades rurais, que tem por objetivo compor uma base de dados integrada de informações das posses rurais com as informações ambientais [Brasil 2012]. Os dados do CAR podem ser acessados por meio do Sistema Nacional de Cadastro Ambiental Rural (SICAR), de responsabilidade do Serviço Florestal Brasileiro (SFB). A realização do cadastro é de responsabilidade do proprietário ou arrendatário rural do imóvel.

A autodeclaração teve o intuito de facilitar e agilizar a inscrição dos produtores [Oliveira and Oliveira 2019], mas associados a essa estratégia também vieram os problemas topológicos oriundos de sobreposições geométricas dos dados declarados. Tais

problemas já eram previstos, com a expectativa de resolvê-los durante o processo de validação dos cadastros, de responsabilidade das unidades federativas. Entretanto, mesmo após 10 anos desde a criação da lei que institui o CAR, até agosto de 2022 foram contabilizados mais de 6,7 milhões de cadastros, dos quais cerca de 1,6 milhões (23%) tiveram algum tipo de análise e apenas 32,5 mil cadastros tiveram a análise de regularidade ambiental concluída [SFB 2022].

Como resultado da autodeclaração e também do lento processo de validação, a base do CAR segue apresentando inúmeros problemas de sobreposições entre imóveis, resultando em problemas topológicos, o que dificulta e até mesmo desestimula a utilização desses dados em diversas pesquisas e em aplicações em Sistemas de Informação Geográfica (SIG). Sendo assim, enquanto a validação oficial desses dados não ocorre, são necessários mecanismos para remover as sobreposições geométricas e ampliar o uso dessa base de dados pelos usuários. Por essa razão, o presente trabalho apresenta o complemento "Q-CleanCAR", em sua primeira versão, que tem como principal função a remoção das sobreposições entre feições dos imóveis rurais cadastrados, baseado na metodologia proposta em [Matias 2022], que propõe a remoção de sobreposição entre os imóveis a partir de uma hierarquia que prioriza os limites dos imóveis de menor área em relação aos maiores.

2. Implementação do complemento

A implementação do complemento "Q-CleanCAR" foi realizada na versão 3.22.5 do software QGIS, utilizando os seguintes complementos: 1) *Processing Modeler* para escolha e teste de funções a serem utilizadas; 2) *Plugin Reloader* para auxiliar no teste das funcionalidades do plugin; 3) *Plugin Builder* que fornece um modelo a partir do qual foi criado o complemento; 4) *Qt Design with QGIS 3.22.5 custom widgets* para construção do layout do complemento. Foi utilizada a linguagem *Python* e o software *Notepad++* para edição dos códigos. O arquivo extensão .zip para instalação do complemento para o QGIS está disponível no site do Laboratório de investigação em Sistemas Socioambientais (LISS) do Instituto Nacional de Pesquisas Espaciais (INPE) e pode ser acessado em: <https://www.lissinpe.com.br/codigos2022>

3. Layout e funcionalidades do complemento

O complemento "Q-CleanCAR" realiza o cálculo semi-automático da sobreposição de áreas entre os imóveis e também permite a realização de edições no dado, disponibilizando desde operações mais básicas que não alteram os limites dos imóveis até operações mais complexas que alteram os limites e arranjos topológicos originais. Para cada opção de edição selecionada é recalculada a área total dos imóveis, sendo possível acompanhar quais etapas e operações impactam mais ou menos o dado com sobreposição original.

Os procedimentos executados no complemento "Q-CleanCAR" são agrupados em 2 eixos: 1) entrada e saída de dados; e, 2) configurações de edição.

3.1. Entrada e saída de dados

A entrada de dados requer a seleção do arquivo *shapefile* referente aos limites da área dos imóveis rurais (obtidos no SICAR) e que sejam informado o tamanho do módulo fiscal e a fração mínima de parcelamento, em hectares, do município (Figura 1a). Nessa versão, só é possível a análise dos imóveis rurais de um único município (Figura 1b).

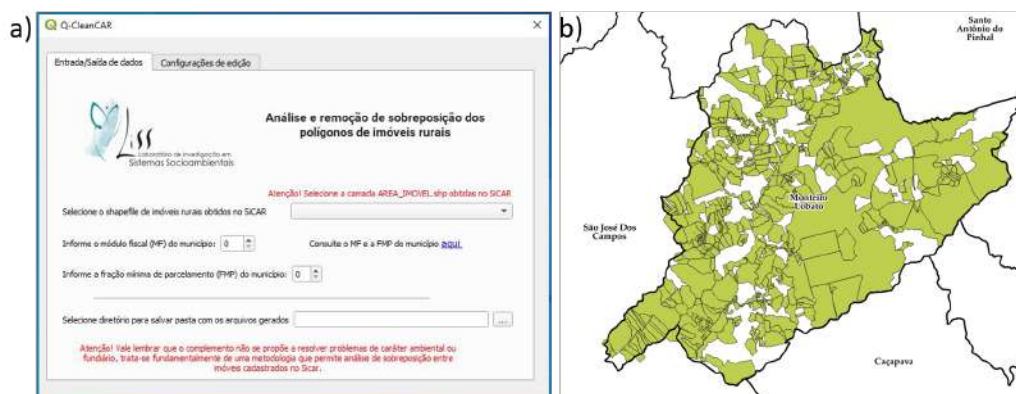


Figura 1. a) Layout da aba de entrada/saída dos dados. b) Feições de imóveis rurais cadastrados no SICAR: exemplo do município de Monteiro Lobato - SP.

Os arquivos gerados são salvos dentro de uma pasta nomeada automaticamente de saídaAAAAMDDHHMMSS (AAAA = ano, MM = mês, DD = dia, HH = hora, MM = minuto e SS = segundo), referente à data e horário da criação dos arquivos dentro do diretório selecionado na primeira aba do complemento.

3.2. Configurações de edição

A segunda aba do complemento (Figura 2) permite a seleção das configurações de edição disponíveis em 6 etapas:

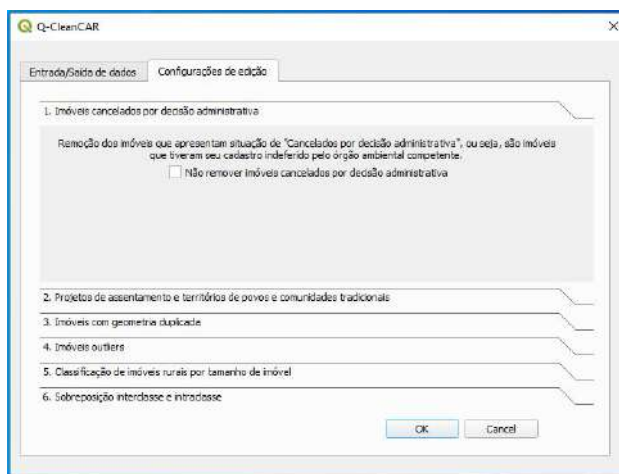


Figura 2. Layout da aba de configurações de edição.

1. Imóveis cancelados por decisão administrativa (Figura 2): Os imóveis cancelados por decisão administrativa são imóveis que tiveram seu cadastro indeferido por órgão ambiental competente. Por padrão, o complemento remove os imóveis cancelados por decisão administrativa, entretanto, nessa guia o usuário poderá optar pela não remoção. Optando-se pela remoção desses imóveis, os polígonos são deletados do ar-

quivo principal e exportados automaticamente como um novo *shapefile* com as feições dos imóveis cancelados.

2. Projetos de assentamento e territórios de povos e comunidades tradicionais (Figura 3): No SICAR há 3 tipos de opção de cadastro: “Imóvel Rural - IRU”, “Imóvel Rural de Povos e Comunidades Tradicionais - PCT” e “Imóvel Rural de Assentamentos da Reforma Agrária - AST”. Por padrão, o complemento mantém os três tipos de cadastro, porém nessa nessa guia o usuário poderá optar por sua remoção, caso deseje realizar as análises somente com os imóveis rurais (IRU). Optando-se pela remoção desses imóveis, os polígonos são deletados do arquivo principal e exportados automaticamente como um novo *shapefile* com as feições dos imóveis AST e PCT.

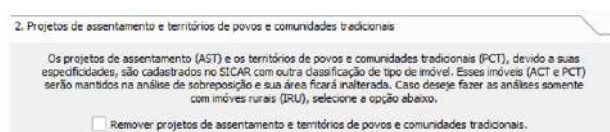


Figura 3. Etapa 2: Projetos de assentamento e territórios de povos e comunidades tradicionais.

3. Imóveis com geometria duplicada (Figura 4): Imóveis com geometria duplicada correspondem aos polígonos de mesma delimitação que foram registrados mais de uma vez no sistema de cadastro. Nessa guia o usuário poderá optar pela remoção de geometrias duplicadas, na qual apenas uma das feições é mantida. A opção padrão é de não realização dessa operação, para que ela ocorra deve-se optar pela remoção.

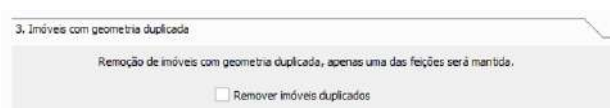


Figura 4. Etapa 3: Imóveis com geometria duplicada.

4. Imóveis outliers (Figura 5): Imóveis rurais *outliers* são aqueles que apresentam área discrepante do conjunto de imóveis do município. Essa opção foi disponibilizada no complemento considerando a metodologia empregada por [Franco 2018], que estabelece um valor de corte a partir do qual um imóvel será considerado ou não um imóvel *outliers* (Equação 1).

$$c = \bar{x} + N * \sigma \quad (1)$$

Onde, c = valor de corte; $\bar{x}$ = Média das áreas dos imóveis rurais no município; N = Número inteiro a ser multiplicado pelo desvio padrão; σ = Desvio padrão das áreas dos imóveis rurais no município.

Por padrão essa operação de remoção de *outliers* não é realizada, porém o usuário poderá optar pela remoção desses imóveis, necessitando apenas informar o valor de N . Optando-se pela remoção desses imóveis, os polígonos são deletados do arquivo principal e exportados automaticamente como um novo *shapefile* com as feições dos imóveis *outliers*.

4. Imóveis outliers

Remoção de imóveis com área discrepante do conjunto de imóveis do município, com base em sua média e desvio padrão. Esses imóveis, potencialmente cadastrados erroneamente, podem sobrepor outros imóveis.

☐ Remover outliers

Selecione o valor para remoção de outliers: média da área dos imóveis + * pelo desvio padrão

Quanto menor o valor de corte, maior a quantidade de imóveis que serão classificados como outliers.

Cuidado! Alguns imóveis que são apenas imóveis grandes podem ser excluídos mesmo não sobrepondo imóveis vizinhos.

Figura 5. Etapa 4: Imóveis outliers.

5. **Classificação de imóveis rurais por tamanho de imóveis (Figura 6):** Os imóveis rurais são classificados por tamanho [Brasil 1993, Brasil 2017], considerando o módulo fiscal (MF) e a fração mínima de parcelamento (FMP). O complemento invariavelmente realiza a classificação dos imóveis rurais (IRU).

5. Classificação de imóveis rurais por tamanho de imóvel

Baseado na fração mínima de parcelamento e no módulo fiscal do município, os imóveis serão classificados em minifúndio, pequeno, médio ou grande.

Os projetos de assentamento e os territórios de povos e comunidades tradicionais não serão classificados.

Classificação dos imóveis	
Minifúndio	< FMP
Pequeno	FMP até 4 MF
Médio	4 MF até 15 MF
Grande	> 15 MF

Figura 6. Etapa 5: Classificação de imóveis rurais por tamanho de imóveis.

6. **Sobreposição interclasse e intraclasse (Figura 7):** A remoção da sobreposição interclasse se dá por meio da diferença espacial entre as 6 classes de imóveis rurais (imóveis AST e PCT, IRU Minifúndio, IRU Pequeno, IRU Médio e IRU Grande), na qual os imóveis AST e PCT cadastrados no SICAR mantêm-se inalterados e são considerados como base para a remoção de sobreposição hierárquica entre as demais classes de imóveis.

Os procedimentos de remoção de sobreposição interclasse geram modificações nos limites dos polígonos dos imóveis. O complemento identifica a existência de áreas sobrepostas entre classes distintas de imóveis e atribui essa área a apenas uma das classes, a partir de uma hierarquia pré-estabelecida que remove a sobreposição entre os imóveis priorizando os limites dos imóveis de menor área em relação aos maiores, resumida em 4 níveis:

- Nível 1: no caso de haver imóveis de projetos de assentamentos e de territórios de povos e comunidades tradicionais que sejam sobrepostos por imóveis de outras classes, os limites dos polígonos AST e PCT são mantidos integralmente, sendo a interseção entre eles removida dos polígonos das demais classes.
- Nível 2: após passar por edições de Nível 1, caso os minifúndios sejam sobrepostos por imóveis pequenos, médios ou grandes, a área de interseção entre eles é mantida nos imóveis minifúndios e removida dos polígonos dos imóveis de maior tamanho.
- Nível 3: após passar por edições de nível 1 e 2, caso os pequenos sejam sobrepostos por imóveis médios ou grandes, a área de interseção entre eles é mantida nos imóveis pequenos e removida dos polígonos dos imóveis médios e grandes.
- Nível 4: após passar por edições de nível 1, 2 e 3, caso os médios sejam sobrepostos por imóveis grandes, a área de interseção entre eles é mantida no imóvel médio e removida dos polígonos dos imóveis grandes.

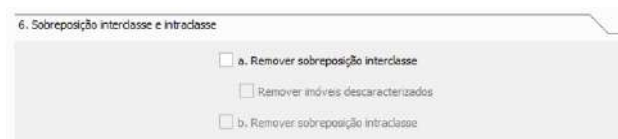


Figura 7. Etapa 6: Sobreposição interclasse e intraclasses.

Caso o usuário selecione a remoção de sobreposição interclasse, pode ocorrer a descaracterização dos polígonos dos imóveis, devido a perda de área e alteração de seu perímetro devido a operação de diferença espacial. Há a opção do usuário selecionar a remoção desses imóveis descaracterizados, os polígonos são deletados do arquivo principal e exportados automaticamente como um novo *shapefile* com as feições dos imóveis descaracterizados.

O processo de remoção de sobreposição intraclasses é realizado através da operação de auto-intersecção de polígonos da camada dos imóveis. O que essa operação faz é delimitar um polígono novo nas áreas sobrepostas de mesma classe, o qual receberá na tabela de atributos a informação de todos os códigos de imóveis envolvidos naquela sobreposição.

Em suma, caso o usuário opte pela realização das duas edições de remoção de sobreposição, interclasse e intraclasses, ele terá ao final um arquivo com os polígonos dos imóveis sem problemas topológicos provocados pela sobreposição de áreas.

4. Considerações finais

O complemento “Q-CleanCAR” busca facilitar o uso do dado de imóveis rurais do CAR em diversas aplicações que utilizem sistemas de informações geográficas. A ideia central é possibilitar para os usuários do dado do CAR um instrumento para compreender a ocorrência de sobreposição geométrica e de problemas topológicos que esse dado pode conter. Para versões futuras do complemento, outras funcionalidades podem ser implementadas, como o acesso automático aos dados do SICAR, bem como rodar o processo em lote, para diversos municípios.

Referências

- Brasil (1993). Lei nº 8.629, de 25 de fevereiro de 1993.
- Brasil (2012). Lei nº 12.651, de 25 de maio de 2012.
- Brasil (2017). Lei nº 13.465, de 11 de julho de 2017.
- Franco, U. E. G. (2018). Avaliação das sobreposições geométricas de imóveis e reservas legais do Cadastro Ambiental Rural (CAR) na RIDE-DF.
- Matias, M. R. (2022). Cartografias da agricultura urbana: Contribuições ao planejamento territorial na região metropolitana do vale do Paraíba e litoral norte.
- Oliveira, L. and Oliveira, F. H. (2019). Verificação da integração entre o Cadastro Ambiental Rural (CAR) e o Cadastro Nacional de Imóveis Rurais (CNIR) sob a ótica do cadastro territorial multifinalitário. *Geosul*, 34(70):339–357.
- SFB (2022). *Regularização Ambiental, Boletim Informativo: Dados declarados até 01 de agosto de 2022.*

Innovative Tools for TerraHidro: An Intuitive GIS User Interface

Thiago L. M. Barreto¹, Sergio Rosim¹

¹Divisão de Observação da Terra e Geoinformática – Instituto Nacional de Pesquisas Espaciais (INPE)
São José dos Campos SP Brazil

{thiago.luiz, sergio.rosim}@inpe.br

Abstract.

This article shows the development and implementation of the TerraHidro system, interface version, in Linux, with innovations in relation to the traditional interfaces of geographic information systems. The objective is to reduce user interaction with the system, allowing several functions to be performed with just one interaction, with intermediate functions being performed automatically. Therefore, it was necessary to create a metadata structure to associate each data with its type. In addition, a tool was implemented to save the execution history, under user command. A text file will be created. It will work with a TerraHidro function script in the command line version.

1. Motivation

Traditionally, GIS (Geographic Information Systems) contains complex interfaces with the characteristic of demanding a large number of interactions from the user, who is normally an expert. Every new step requires an action from the operator informing their respective parameters and necessary file paths, which may be problematic for a continuous workflow environment. This is, the person needs to be fully dedicated for every new step avoiding the process to stop.

This paper presents the performed work to change the amount of operator interactions with the TerraHidro system interface, which is a GIS developed for distributed hydrological modeling tasks. It was achieved an important improvement on the user interface by reducing the operator dependency on the system, which is now able to execute several tasks in sequence by just configuring one interface instead of many. That indicates less time spent for setting up the interface and reduces the chance of introducing human errors. This was reached by implementing a generalized metadata schema mapped for the workflow and subject to further improvements. In addition, it is possible to save the processing steps and parameters into a history file for each executed interface, enabling the operator to have an exact batch script version of his work.

2. Related Works

It has been notable how Human Computer Interfaces (HCI) have recently evolved, especially those from mobile devices [2] where a certain level of adaptation was achieved between users and their environment (i.e. automatic adjustment of screen brightness based on surrounding light levels). Some other systems were automated in

order to find groups based on their roles with a view to software engineering [4] and designing [3]. Thanks to studies with focus on the interface design and its layout components [5] we are able to understand better how users' performance interacts with respect to agronomy [6] (e.g. font size/color, element displacement) and software engineering driven development/tests [7, 8, 9].

When it comes to GIS interfaces we understand how complex this issue is since it normally demands highly skilled professionals for operating it as a result of the lack in design standards [1]. This happens due to complexity of interoperability [10, 11] in terms of: data (file formats, I/O), functional (CPU/GPU processing, memory), and device (web-based, desktop-based).

Recently, there were suggested different approaches for designing interfaces based on the result of quality attributes (hedonic or pragmatic) from users evaluation questionnaires [12, 13, 14, 15]. This not only supports more efficient software engineering prototyping cycles but also enables the evaluation between two or more distinct interfaces.

This paper aims to present a more intuitive GIS interface by an approach that drives the user to automatically set up the correct parameters and file paths based on previous processing stored metadata.

3. The TerraHidro Concepts

In general, systems have the property of approaching users with little computational knowledge, called laymen, of complex functionalities. Typically, this type of user is focused on producing results that are often repetitive. The characteristics of these GIS are the strong interaction between user and system. What is proposed in this work is the optimization of the time spent by the user by reducing the necessary interactions with TerraHidro GIS. For this, the TerraHidro interface was reprogrammed. A metadata schema was developed to initially contain the type of each data in the TerraHidro scope. Finally, the history of the functions executed, under user command, was also implemented, which is not common in systems with interfaces.

The user interface for TerraHidro has been designed and developed under TerraLib/TerraView framework, which is an open source library for Desktop customized GIS applications written and compiled in C/C++ under the GNU Lesser General Public License (LGPL). All new functionalities can be found as a shared library plugin compiled and installed into Terraview allowing new interfaces to be accessed through the main menu.

In this section, we present in detail the newly designed interfaces for hydrological modeling, the metadata implementation which allowed an important improvement on user experience, and finally the usability of storing every system call into batch scripts.

3.1. Interfaces

Each interface was designed and implemented with respect to the general workflow diagram presented in Figure 1 as well as its respective relationship in classes containing their inputs and outputs. The process starts with a given Digital Elevation Model (DEM) of a scene being pre-processed for removing flat areas and pits (REMOVE_PITS).

Then, each pixel is computed with respect to its local direction flow inside a 8 neighbor (D8) window. This allows the system to compute the accumulation of the flow (D8) and to extract the river streams by applying a cut-off threshold classification (DRAINING). Finally, every individual river stream is calculated (SEGMENTS) in order to obtain its respective river basin (MINIBASINS).

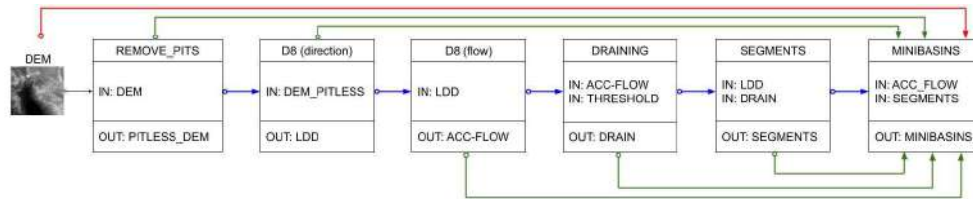


Figure 1. Workflow diagram of TerraHidro main functions.

This workflow can be executed in two different ways. Step by step, when the user needs to command the execution of each function, in the following sequence: REMOVE_PITS, D8, D8CA, DRAINING, SEGMENTS e MINIBASINS. It's shown in the Figure 1 by the blue line sequences. Another way is the direct execution of the desired function, MINIBASINS in this case, from the existing input data. In this case, the user will only have to interact once, as the intermediate functions are being performed automatically. It's presented by the red line.

In this figure, the user starts by giving the original altimetry grid as input data. The possibilities are shown Once activated, the MINIBASINS function will cause the system to execute REMOVE_PITS, D8, D8CA, DRAINING, and SEGMENTS functions automatically, generating their respective data files. The concept will be the same if the initial data is the altimetry grid already with the flat areas and local minima extracted, or the grid of local flows, or the grid of accumulated flows, or the drainage grid, or even the grid of segments of drainage. The other possibilities can be seen in the green lines at the same figure.

3.2. Metadata

The type of each data file is critical for this interface design to work. The data type will indicate which functions must be performed, according to the function performed by the user. For example, the user has an altimetry grid as input and aims to execute the DRAINING function. The data type informs if the grid is of original altimetry (DEM) or if it has already been processed to eliminate flat areas and local minima. Therefore, each raster file will have at least one metadata that is its type. All metadata information is stored in a .json data structure, described below.

In the literature it is possible to find good examples of metadata being used for geoinformation systems. Among them, we can mention those derived from ISO institutional standards or free code from the W3C (World Wide Web Consortium) [16,17], and which supported the adoption of the standard presented below. After numerous discussions with different stakeholders, we came up with a metadata model to be manipulated by the system in .json file format. Figure 2 presents the final data structure, as well as their respective fields and description. The .json data structure model was chosen due to its wide adherence among today's main development tools, and can be easily integrated with them, in addition to being an open source file standard.

Metadata	Description
"id": "000000"	A unique string key containing the file ID
"predecessors": ["000", "111"]	A set of file IDs corresponding to all predecessors
"pullpath_filename": "path"	A string field containing the full path of the file
"process_type": "type"	A string field to carry on which step of workflow was responsible to generate that file specifically
"threshold": { "absolute": "value", "percentual": "value" }	A struct for storing which threshold was used for binarization of river streams
"extras": []	A free space for user custom metadata storage

Figure 2. Metadata structure and description.

It was also discussed possible limitations of the methodology such as the number of attributes, generalization of parameters and inclusion/reading of attributes by the user. All of them were solved by implementing a metadata class called “extras” as described in Figure 3. Within this data structure, it is possible to define metadata different from what was initially planned.

```

"extras": [
  {
    "pointName": "point01",
    "pointDesc": "gps point close to river",
    "XCoord": "100",
    "YCoord": "200",
    "Lat": "-23",
    "Lon": "-12"
  },
  {
    "pointName": "point02",
    "pointDesc": "gps point close to town",
    "XCoord": "1000",
    "YCoord": "2000",
    "Lat": "-23.01",
    "Lon": "-12.01"
  }
]

```

Figure 3. A piece of a metadata file generated with the “extras” function to store data related to geographic coordinates.

It is considered a file that is imported into the TerraView geographic viewer, which hosts the TerraHidro plugin. By the first time a TerraHidro function is executed, the user will receive a warning alert that the input file has its unspecified. The user must then associate the appropriate type, ensuring that any data used by TerraHidro has its data type.

3.3. Interface to Batch Script (history of executions)

The third feature we presented on our GIS is the possibility for the user to store the execution history in a .txt file. He defines the starting and ending point of the file and the generated file will contain the functions performed during the execution of TerraHidro interfaces, already in the TerraHidro command lines format.

For instance, considering the workflow executed from SEGMENTS, as shown in the Figure 4, step and having already run PITLESS_DEM as shown in bold in Figure 4, will produce the following batch script accordingly. Note that for each executed step, the filenames are being defined automatically in order to be convenient.


```
./th d8 /home/inpe/data/s13_w040_1arc_v3_PITLESSDEM.tif
/home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD.tif
./th d8ca /home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD.tif
/home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD_ACC-FLOW.tif
./th d8drainage /home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD_ACC-FLOW.tif
/home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD_ACC-FLOW_DRAIN_100_00.tif
12329.30
./th segments /home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD.tif
/home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD_ACC-FLOW_DRAIN_100_00.tif
/home/inpe/data/s13_w040_1arc_v3_PITLESSDEM_LDD_ACC-
FLOW_DRAIN_100_00_SEGMENTS.tif
```

Figure 4. Resulting batch script from storing the execution history based on Figure 1.

4. Application example

This example shows the execution of the DRAINING function from the original altimetry grid. Before, this grid was imported into TerraHidro, receiving the DEM data type as metadata. The user just had to choose this function by selecting the altimetry grid and pressing the "execute" button. Figure 5 shows the original grid, the interface used and the drainage result, in vector format.

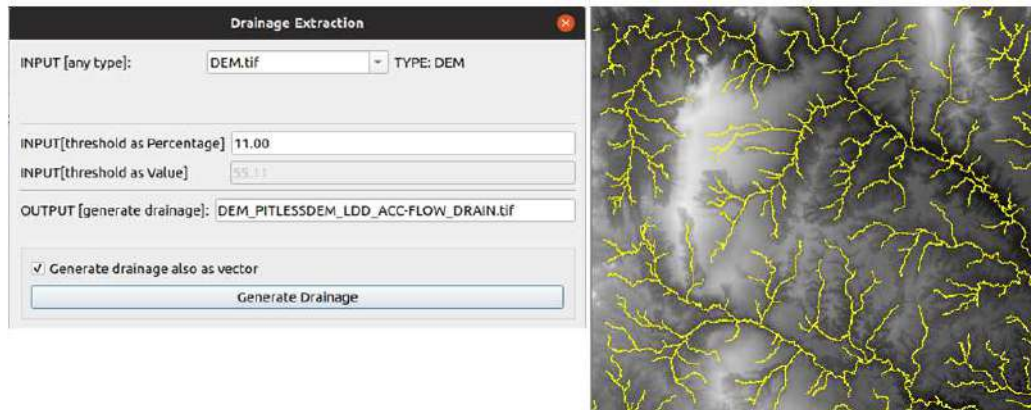


Figure 5: interface for drainage extraction and altimetry grid and drainage extracted.

5. Conclusions

The objective of this work was the development of the TerraHidro system, interface version, which could optimize the user's time in their interactions with this system. In addition to the step-by-step processing, the user can skip some steps that are automatically executed by the system, which, recognizing the initial data by its type, knows which functions must be executed. For that, a metadata structure was created to support the type of each data imported or created in the system, also allowing the user to create new metadata values. Another differential of this interface is the possibility of saving the execution log in the format of TerraHidro's own command lines.

6. Acknowledgments

The authors would like to thank the Brazilian National Research Council (CNPq) for the research funding spent fellowship of first author.

References

- [1] GOODCHILD, Michael F. Spatial thinking and the GIS user interface. **Procedia-Social and Behavioral Sciences**, v. 21, p. 3-9, 2011.
- [2] HUSSAIN, Jamil et al. Model-based adaptive user interface based on context and user experience evaluation. **Journal on Multimodal User Interfaces**, v. 12, n. 1, p. 1-16, 2018.
- [3] ZUBKOVA, T.; TAGIROVA, L. Intelligent user interface design of application programs. In: **Journal of Physics: Conference Series**. IOP Publishing, 2019. p. 012026.
- [4] MARTINEZ, Jabier et al. Variability management and assessment for user interface design. In: **Human Centered Software Product Lines**. Springer, Cham, 2017. p. 81-106.
- [5] SHAO, Junkai et al. How dynamic information layout in GIS interface affects users' search performance: integrating visual motion cognition into map information design. **Behaviour & Information Technology**, p. 1-18, 2022.
- [6] WANG, Siyi et al. Interface Design of GIS System Based on Visual Complexity. In: **International Conference on Applied Human Factors and Ergonomics**. Springer, Cham, 2019. p. 713-722.
- [7] SILVA, Thiago Rocha; WINCKLER, Marco; TRÆTTEBERG, Hallvard. Extending Behavior-Driven Development for Assessing User Interface Design Artifacts. In: **The 31st International Conference on Software Engineering & Knowledge Engineering (SEKE 2019)**. 2019.
- [8] SAMRGANDI, Najwa. User Interface Design & Evaluation of Mobile Applications. **International Journal of Computer Science & Network Security**, v. 21, n. 1, p. 55-63, 2021.
- [9] KURIA, Ezekiel; KIMANI, Stephen; MINDILA, Agnes. A framework for Web GIS development: a review. **International Journal of Computer Applications**, v. 178, n. 16, p. 6-10, 2019.
- [10] KUPER, Paul Vincent. Design of a Generic Mobile GIS for Professional Users. **ISPRS International Journal of Geo-Information**, v. 7, n. 11, p. 422, 2018.
- [11] AGRAWAL, Sonam; GUPTA, R. D. Web GIS and its architecture: a review. **Arabian Journal of Geosciences**, v. 10, n. 23, p. 1-13, 2017.
- [12] KAMMERHOFER, David; SCHOLZ, Johannes. An approach to decompose and evaluate a complex GIS-application design to a simple, lightweight, user-centered app-based design using user experience evaluation. **ISPRS International Journal of Geo-Information**, v. 9, n. 9, p. 505, 2020.
- [13] PYKÄLÄINEN, Tuukka et al. Making location data visual—Simplifying an online municipality-GIS presentation editor through user interface design. 2019.
- [14] CHARFI, Selem; EZZEDINE, Houcine; KOLSKI, Christophe. RITA: a useR Interface evaluation framework. **J. Univers. Comput. Sci.**, v. 21, n. 4, p. 526-560, 2015.
- [15] ROT, Artur; KUTERA, Robert; GRYNCEWICZ, Wiesława. Design and Assessment of User Interface Optimized for Elderly People. A Case Study of Actgo-Gate Platform. In: **ICT4AgeingWell**. 2017. p. 157-163.
- [16] DO PRADO, Bruno Rodrigues et al. Padrões para metadados geográficos digitais: modelo ISO 19115: 2003 e modelo FGDC. *Revista Brasileira de Cartografia*, v. 62, 2010.
- [17] ALBERTONI, Riccardo; DE MARTINO, Monica; QUARATI, Alfonso. Documenting context-based quality assessment of controlled vocabularies. *IEEE Transactions on Emerging Topics in Computing*, v. 9, n. 1, p. 144-160, 2018.

Implementação de Ferramentas em *Software* Livre para Avaliação da Acurácia Posicional em Produtos Cartográficos

Isaac Antunes Miranda Junior¹, Afonso de Paula dos Santos¹, Marconi Martins Cunha¹, Lúgia da Silva Barbosa¹

¹Departamento de Engenharia Civil – Universidade Federal de Viçosa (UFV)
Av. Peter Henry Rolfs, s/n - Campus Universitário – 36570-900 – Viçosa – MG – Brasil

{isaac.junior, afonso.santos, marconi.cunha, ligia.barbosa}@ufv.br

Abstract. *In Brazil, there are standards for evaluating one of the most important elements of cartographic quality control, which is positional accuracy. Among them, the following stand out: Decree No. 89,817 of 1984 and the Technical Specification for Quality Control of Geospatial Data (ET-CQDG) of the National Spatial Data Infrastructure (INDE); the INCRA Normative Instruction of 2018 and NBR 13,133 of 1994. However, there is a scarcity of cartographic quality control tools implemented in free-form Geographic Information System (GIS) software. Thus, this work presents the development of a plugin in QGIS for the assessment of positional accuracy of cartographic data and products based on current regulations in the country.*

Resumo. *No Brasil existem normas visando a avaliação de um dos elementos mais importantes do controle de qualidade cartográfica, que é a acurácia posicional. Dentre elas, destacam-se: Decreto nº 89.817 de 1984 e a Especificação Técnica de Controle de Qualidade de Dados Geoespaciais (ET-CQDG) da Infraestrutura Nacional de Dados Espaciais (INDE); a Instrução Normativa do INCRA de 2018 e a NBR 13.133 de 1994. Entretanto, há uma escassez de ferramentas de controle de qualidade cartográfica implementadas em softwares de Sistema de Informação Geográfica (SIG) de forma livre. Assim, este trabalho apresenta o desenvolvimento de um plugin no QGIS para a avaliação da acurácia posicional de dados e produtos cartográficos com base nas normas vigentes no país.*

1. Introdução

A Cartografia sempre foi e será uma área de segurança estratégica. Segundo Brotton (2014), desde o século V a.C. a elite grega já fazia uso de mapas para orientar suas conquistas, como realizado por Alexandre, o Grande. Posteriormente, na idade média, só os reis detinham as cartas, sendo o conhecimento do território uma forma de poder. Este acesso restrito à cartografia continuou até o início do século XX. Contudo, a partir da segunda metade deste século, os mapas e as cartas se tornaram mais acessíveis.

A partir dos anos 2000, com a popularização da *internet*, *smartphones* e uma dinâmica social eminentemente digital, o uso da Cartografia foi incorporado ao cotidiano da grande maioria da população. Neste contexto, segundo Santos (2015), cada vez mais a informação passou a necessitar da componente espacial. O simples ato de ir ao trabalho, viajar ou até mesmo solicitar um táxi ou pedir um *delivery* requer a utilização de mapas. Outras áreas de tecnologias de produção, como infraestrutura, obras de engenharia, agricultura de precisão, telecomunicações e o setor de serviços dependem fortemente do

conhecimento e representação da superfície terrestre. Mendonça e Sluter (2011) reforçam esta ideia, ao apontar que a produção cartográfica é onerosa, mas de importância estratégica para infraestrutura e segurança de um país.

Entretanto, com o crescimento de produtores e usuários das informações espaciais, uma das necessidades que surge é atestar a qualidade dos produtos gerados [Lunardi, Penha e Cerqueira 2012]. Nesse contexto, o Controle de Qualidade Cartográfica (CQC) avalia a qualidade de um produto cartográfico para que se possa ter confiabilidade nele, de modo que o usuário o utilize de maneira adequada para o fim a que se destina. Essa avaliação é realizada sobre os elementos de qualidade dos produtos. Segundo ISO (2013), os elementos básicos da qualidade a serem observados e avaliados nos produtos cartográficos são: completude, consistência lógica, acurácia posicional, acurácia temporal, acurácia temática e usabilidade.

Dentre esses elementos, a acurácia posicional é um dos mais estudados e de maior importância [Santos 2015]. Devido a isso, a acurácia posicional é o elemento do CQC mais utilizado para verificar a qualidade dos dados espaciais [Jakobsson e Vauglin 2002]. Tradicionalmente, essa avaliação da acurácia posicional é realizada por meio das discrepâncias posicionais entre a localização de uma entidade espacial no produto geoespacial avaliado e sua localização de referência, de melhor acurácia [ANSI 1998]. Sendo que, geralmente, essa verificação é realizada por meio de técnicas baseadas em pontos [Ariza-López, Rodríguez-Avi e Alba-Fernández 2018].

O processo de controle de qualidade posicional normalmente é realizado tomando-se um padrão ou norma como referência. No Brasil, as principais normas que abordam a avaliação acurácia posicional de produtos cartográficos são: (i) o Decreto nº 89817 de 1984 e a Especificação Técnica de Controle de Qualidade de Dados Geoespaciais (ET-CQDG) da Infraestrutura Nacional de Dados Espaciais (INDE); (ii) a Instrução Normativa INCRA/DF/n.02 de 2018; e (iii) a NBR 13.133 (versão 1994).

Para que a implementação destas normas e padrões seja efetiva e os usuários possam ter confiabilidade no produto cartográfico avaliado, são necessárias ferramentas acessíveis em *softwares*, principalmente de Sistemas de Informações Geográficas (SIG). Atualmente, a automação do controle de qualidade cartográfica é incipiente em todo o mundo. No Brasil, se tem conhecimento apenas do *software* GeoPEC, que é uma aplicação computacional *desktop*. No QGIS, o SIG em *software* livre mais utilizado no mundo, não há nenhuma ferramenta que possibilite avaliar a acurácia posicional em produtos cartográficos com o uso dos padrões e normas brasileiras. É fácil perceber que uma ferramenta desenvolvida para *software* livre possui um potencial de alcance maior, por atingir um número de usuários maior.

Diante disso, o objetivo deste trabalho foi criar ferramentas, por meio de um *plugin* no *software* QGIS para avaliação da acurácia posicional em produtos cartográficos, que utilizem as principais normas e padrões sobre a avaliação acurácia posicional vigentes no Brasil.

2. Desenvolvimento de um *plugin*

Um *plugin* tem como objetivo adicionar novas funcionalidades a um determinado *software*. Ele pode ser definido como um programa a parte e que não pode ser executado individualmente, pois utiliza o *software* anfitrião para o controlar e ter o acesso aos dados [Fontoura et. al. 2016].

Para a construção do *plugin*, denominado PEC-Avalia, foram utilizados o QT Designer, a linguagem Python e os complementos do QGIS *Plugin Builder* e *Plugin Reloader*. O *Plugin Builder* auxilia na criação da estrutura de arquivos e pasta necessário para o desenvolvimento e o funcionamento de *plugins* no QGIS. Para posterior adição de funcionalidade da ferramenta, utilizou-se a linguagem Python. Já o *Plugin Reloader* atualiza o *plugin* a cada alteração no seu código base. Para o desenvolvimento da interface foi utilizado o QT Designer. A Figura 1 apresenta a interface do *plugin* PEC-Avalia.

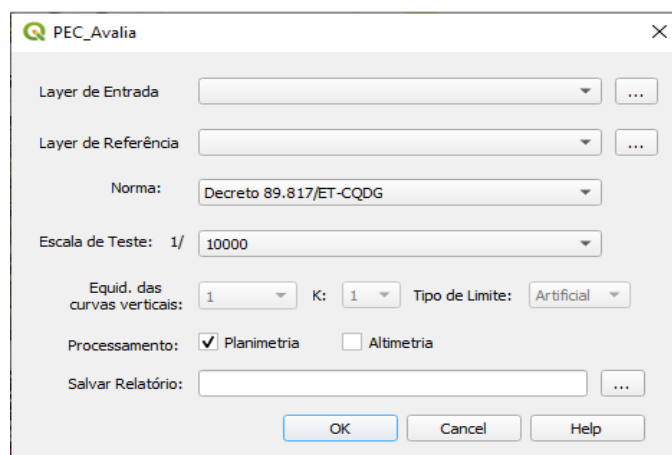


Figura 1. Interface gráfica do PEC-Avalia

Após a instalação, o *plugin* pode ser acessado no QGIS pelo menu complementos. O algoritmo do *plugin* PEC-Avalia pode ser observado na Figura 2.

Todo o código foi escrito em Python, tendo como parâmetros iniciais as *layers* (camadas) de entrada em formato *shapefile*, que correspondem ao produto cartográfico de teste e os dados de referência. Após a entrada destes parâmetros, o usuário deve escolher a norma desejada. Após realizar esta escolha, novos parâmetros são habilitados de acordo com a norma em questão.

Ao selecionar a Instrução Normativa do INCRA, é realizada uma checagem para verificar se a amostra é composta por 20 ou mais pontos, como determina a norma. Se houver menos de 20 pontos, é apresentada uma mensagem de erro e o processamento é cancelado. Caso contrário, deve ser informado o tipo de limite: artificial, natural ou inacessível. O processamento se inicia com a realização do teste de Shapiro-Wilk (95%) para avaliação da normalidade. Se os dados não apresentarem normalidade, o produto avaliado é reprovado. Caso contrário, é realizado o teste t de *Student* para verificar a presença de tendência nos dados. Se houver tendência, o dado é reprovado. Senão, é avaliado se 100% das discrepâncias posicionais são menores ou iguais à precisão posicional correspondente a cada tipo de limite. Se isto ocorrer, o produto avaliado é aprovado. Por outro lado, se esta porcentagem foi inferior a 100%, o produto cartográfico avaliado é reprovado.

Caso o usuário opte por avaliar um produto segundo a NBR 13.133 (versão 1994), também é realizada uma verificação inicial, relacionada à geometria dos dados. Ao selecionar o tipo de processamento como planimetria, as feições de entrada devem ser do tipo linha. Por outro lado, se o usuário optar por analisar a altimetria, esta camada deve ser do tipo ponto. Se isto não ocorrer, é exibida uma mensagem de erro e o processamento é interrompido. Caso contrário, dependendo da escolha realizada, é solicitado ao usuário

informar a escala de teste ou a equidistância vertical das curvas de nível. Em seguida, deve ser fornecido o valor de K, um coeficiente relativo à classe do equipamento utilizado no levantamento topográfico. O processamento desta norma tem início com o cálculo das discrepâncias das distâncias e/ou altimétricas, dependendo da opção selecionada pelo usuário. Em seguida, é calculado o “desvio-padrão” das distâncias e/ou altimétrico, que é comparado com o “desvio-padrão admissível” definido pela norma. Já as discrepâncias são comparadas com o “padrão de exatidão” da norma. Nesta comparação, duas condições devem ser satisfeitas: 90% das discrepâncias devem ser menores que o “padrão de exatidão”, planimétrico ou altimétrico, e o “desvio-padrão” calculado deve ser menor que o “desvio-padrão admissível”. Se estas duas condições forem satisfeitas, o produto é aprovado. Caso contrário, ele é reprovado.

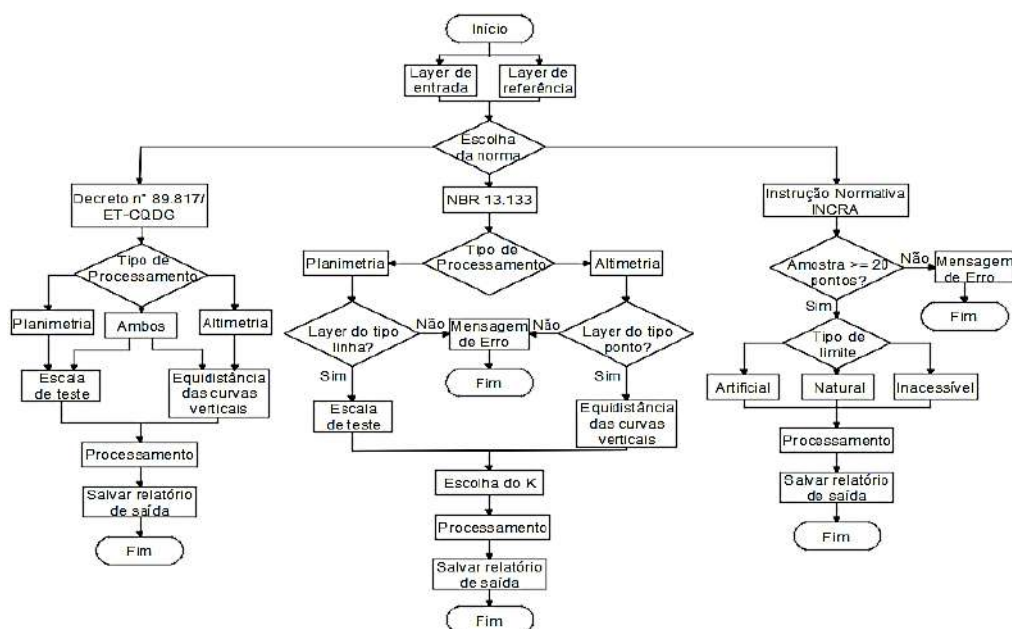


Figura 2. Algoritmo do *plugin* PEC-Avalia

Por fim, o usuário pode optar por avaliar seu produto cartográfico segundo o Decreto nº 89.817/ET-CQDG. Assim como no caso anterior, nesta opção inicialmente deve ser selecionado o tipo de processamento: planimetria, altimetria ou ambos, todos os casos devem ter seus dados de entrada do tipo ponto. Dependendo do tipo de análise selecionada pelo usuário, este deverá fornecer a escala de teste e/ou a equidistância vertical das curvas de nível. O processamento do Decreto nº 89.817/ET-CQDG se inicia com o cálculo das discrepâncias posicionais. Com os valores da escala de teste e/ou equidistância das curvas de nível são calculadas as tolerâncias Padrão de Exatidão Cartográfica (PEC) e Erro-Padrão (EP). Com base nestas tolerâncias, dois critérios devem ser atendidos: 90% das discrepâncias (planimétricas e/ou altimétricas) devem ser menores que o valor do PEC e o RMS (*Root Mean Square*) destas discrepâncias deve ser menor que o EP. Esta avaliação se inicia na classe A. Se o produto for reprovado nesta classe, então ele é avaliado segundo a classe B, e assim sucessivamente. Se o produto avaliado não se enquadrar em nenhuma classe, é dado como reprovado para a escala e/ou equidistância vertical das curvas de nível.

Por fim é gerado um relatório detalhado, em formato .txt, onde, inicialmente, apresenta-se as informações inseridas pelo usuário. Posteriormente, o resultado é detalhado como: estatísticas da média, desvio-padrão, variância, RMS, valores máximos e mínimos das discrepâncias, bem como os valores tabelados que são utilizados na norma escolhida e sua devida porcentagem de aceitação. Ao final, as discrepâncias com seus respectivos ID das feições são escritas no relatório. Na Figura 3 é apresentado um relatório gerado pelo *plugin* PEC-Avalia na avaliação da acurácia planimétrica de uma ortomagem Ikonos, avaliada segundo o Decreto n°89.817/ET-CQDG, para a escala de 1/10.000. Nesta figura é possível observar as informações do produto avaliado e o seu resultado (Figura 3a), as estatísticas e as tolerâncias utilizadas (Figura 3b) e as discrepâncias posicionais (Figura 3c).

PROCESSAMENTO ACURÁCIA POSICIONAL - PEC-AVALIA			
Norma utilizada: Decreto 89.817/ET-CQDG			
Processamento: Planimétrico			
Escala planimétrica: 1/10000.0			
Camada teste: ikonos_IMG_23pts_wgs84_utm			
Camada de referência: ikonos_GNSS_23pts_wgs84_mapgeo84			
RESULTADO PLANIMETRIA:			
Classificação pelo Decreto 89.817/ET-CQDG:			
Classe C na escala 1/10000.0			

(a)

ESTATÍSTICAS:			
Número de pontos utilizados: 23			
	dx	dy	d2d
Média	-0.7946	2.2775	3.3970
Desvio Padrão	1.4375	2.9859	2.2361
Variância	2.0665	8.9153	5.0001
RMS	1.6149	3.7033	4.0401
Máximo	1.4674	7.3446	8.7669
Mínimo	-4.7871	-2.7369	0.3288

(b)

Tolerâncias	PEC-PCD Plan	EP Plan
A	2.00	1.70
B	5.00	3.00
C	8.00	5.00
D	10.00	6.00

ET-CQDG	%di(2D) <= PEC-PCD
A	43.48%
B	82.61%
C	95.65%
D	100.00%

DISCREPÂNCIAS			
ID	d1(E)	d1(N)	d1(2D)
23	-0.0527	3.8672	3.8676
12	-0.1701	7.1287	7.1307
11	-1.1814	6.6279	6.7324
13	-1.2370	3.7077	3.9087
16	0.3259	0.0435	0.3288
10	-1.9504	3.2666	3.8045
9	-2.3418	5.5554	6.0288
14	-0.5278	0.7157	0.8893
20	-0.9511	4.6590	4.7551
22	1.4674	2.0922	2.5535
8	0.8787	-2.7369	2.8745
21	-0.2450	0.3491	0.4264
7	-0.4750	1.7049	1.7699
17	-0.4877	4.4057	4.4326
4	-1.2282	-2.0048	2.3511
5	-4.7871	7.3446	8.7669
3	-1.6917	-0.3851	1.7350
15	-1.4686	-0.5473	1.5673
18	-3.3689	0.6715	3.4352
2	1.1345	-1.7991	2.1269
6	-0.6691	3.3063	3.3733
19	0.0676	4.5683	4.5688
1	0.6843	-0.1593	0.7026

(c)

Figura 3. Exemplo do relatório de saída

Para mais detalhes sobre estas normas, pode-se consultar o Decreto n° 89817 de 1984 [BRASIL 1984], a ET-CQDG [DSG 2016], a Instrução Normativa do INCRA [INCRA 2018] e a NBR 13.133 [ABNT, 1994].

O *plugin* desenvolvido foi disponibilizado e pode ser obtido através do link: <<https://github.com/isaacantunes/PEC-Avalia>> ou obtido pelo próprio *software* QGIS, na aba complementos.

3. Conclusões

Neste trabalho foi desenvolvido um *plugin*, para o *software* livre QGIS, para a avaliação da acurácia posicional segundo os critérios da Instrução Normativa do INCRA, da NBR 13.133 ou do Decreto n° 89.817/ET-CQDG. Tal *plugin* foi disponibilizado gratuitamente em um link do *github* e está disponível no repositório oficial do QGIS, visando facilitar

o acesso de todos que necessitem de uma ferramenta simples para auxiliar no processo de avaliação da acurácia posicional, adotando normas existentes no Brasil.

Como recomendação de trabalhos futuros, sugere-se a implementação de normas de outros países, bem como a implementação de ferramentas para a avaliação de outras componentes do controle de qualidade cartográfica, o que irá permitir uma análise completa da qualidade de produtos cartográficos.

Agradecimentos

Os autores agradecem ao CNPq e a Universidade Federal de Viçosa pelo auxílio no projeto de Iniciação Científica no âmbito do edital PIBIC/UFV/2021-2022.

Referências

- ANSI - American National Standards Institute. Spatial Data Transfer Standard Sdts—Part1: Logical Specifications. Ansi Ncits 320-1998, Washington, D.C, 1998.
- Ariza-López, Francisco Javier; Rodríguez-Avi, José; Alba-Fernández, Virtudes. A Positional Quality Control Test Based on Proportions. In: The Annual International Conference on Geographic Information Science. Springer, Cham, P. 349-363, 2018.
- BRASIL. Decreto nº 89.817 de 20 de junho de 1984. Normas Técnicas da Cartografia Nacional. Brasil, 1984.
- Brotton, J. Uma história do mundo em doze mapas. Jerry Brotton; tradução Pedro Maia, 1ª ed. Rio de Janeiro: Zahar, 2014.
- DSG. ET-CQDG: Especificação Técnica para Controle de Qualidade de Dados Geoespaciais. 1 ed. Brasil, 2016.
- INCRA. Norma de Execução INCRA/DF/N.02 de 19 de fevereiro de 2018. Brasil, 2018.
- Fontoura, J. R.; Allasia, D. G.; Froemming, G. H.; Geller, V. G.; Favaretto, J. R. Geoferramenta Para Aquisição Automática De Dados Pluviométricos a Partir Do Qgis. In: 6º Simpósio De Geotecnologias Do Pantanal, 2016, Cuiabá. Anais Do 6º Simpósio De Geotecnologias Do Pantanal, 2016, 2016.
- ISO. ISO 19157: Geographic information - Data quality. International Organization for Standardization, 2013.
- Jakobsson, A.; Vauglin, F. Report of A Questionnaire on Data Quality In National Mapping Agencies. Cerco Working Group on Quality. Comité Européen Des Responsables De Cartographie Officielle, Marne-La-Vallée, 2002.
- Lunardi, O. A.; Penha, A. L. T.; Cerqueira, R. W. O Exército Brasileiro E Os Padrões De Dados Geoespaciais Para A Inde. In: Simpósio Brasileiro De Ciências Geodésicas E Tecnologias Da Geoinformação. 4, 2012, Recife. P. 001 – 008. 2012.
- Mendonça, A. L. A.; Sluter, C. R. Análise Da Relação Entre Ensino E Pesquisa Em Ciências Geodésicas E A Cobertura Do Mapeamento Sistemático No Brasil. Revista Brasileira De Cartografia, V. 63, 2011.
- Santos, A. P. Controle De Qualidade Cartográfica: Metodologias Para Avaliação Da Acurácia Posicional Em Dados Espaciais. Tese (Doutorado Em Engenharia Civil) – Universidade Federal De Viçosa, Viçosa, 188p, 2015.

Sistema de Geração de Dados Faltantes em Séries Temporais Climatológicas usando Triangulação, Aprendizado e Meta-Aprendizado de Máquinas

Pablo de O. Passarini¹, Filipe J. S. Coelho¹, Marconi A. Pereira¹,
Natã G. Silva¹, Telles T. Da Silva³

¹Departamento de Tecnologias - DTECH
Universidade Federal de São João del-Rei (UFSJ)
Campus Alto Paraopeba, MG 443, km 7, Ouro Branco/MG, Brasil.

²Departamento de Física e Matemática - DEFIM
Universidade Federal de São João del-Rei (UFSJ)
Campus Alto Paraopeba, MG 443, km 7, Ouro Branco/MG, Brasil.

pablo4passarini@gmail.com, marconi@ufsj.edu.br

ngoularts@ufsj.edu.br, timoteo@ufsj.edu.br

Abstract. *This work presents an application developed for the generation of missing data in time series, using meteorological data as input parameters. The application uses data from a set of four different cities, containing three climate indicators: maximum temperature, minimum temperature and precipitation. The proposed tool is able not only to complete time series that have data gaps, but also to generate a complete time series for a city that does not have any data. The system uses data triangulation, machine learning and meta-learning algorithms.*

Resumo. *Este trabalho apresenta uma aplicação desenvolvida para a geração de dados ausentes em séries temporais, usando como parâmetros de entrada dados meteorológicos. A aplicação utiliza dados de um conjunto de quatro cidades distintas, contendo três indicadores climáticos: temperatura máxima, temperatura mínima e precipitação. A ferramenta proposta é capaz não só de completar séries temporais que possuam lacunas de dados, como também gerar uma série temporal completa para uma cidade que não possua nenhum dado. O sistema utiliza algoritmos de triangulação de dados, aprendizado e meta-aprendizado de máquinas.*

1. Introdução

Com o avanço tecnológico dos últimos anos foi possível investir em diversas pesquisas científicas que poderiam ter alto potencial de solução para as problemáticas que constantemente surgem no mundo hodierno. Todavia, as pautas que englobam as questões meteorológicas e climáticas não foram exceção durante esse período, visto que são ferramentas essenciais para a formulação de pesquisas e, portanto, possuem um valor imprescindível para o desenvolvimento climatológico no mercado atual.

Neste artigo é apresentada uma aplicação que pode ser utilizada em análises de dados climáticos e, também, para previsões de séries temporais. Esta aplicação pode auxiliar na tomada de decisões em processos conduzidos por empresas, pessoas e instituições.

Para este objetivo foi elaborada uma Interface Gráfica para o Usuário em Python utilizando a biblioteca *TKinter*¹. O sistema foi desenvolvido com o propósito de ser uma interface simples e de fácil utilização, para um usuário com qualquer nível de conhecimento de informática.

Os dados utilizados neste trabalho foram obtidos no site do Instituto Nacional de Meteorologia (INMET)². Este site possui uma base de dados que contém mais de 400 estações meteorológicas, automáticas e convencionais, distribuídas em todo território brasileiro. Neste trabalho escolhemos utilizar três parâmetros que são a temperatura máxima, a temperatura mínima e precipitação total diária.

A aplicação é uma evolução do trabalho presente em [Alves and Pereira 2020]. Neste trabalho o meta-aprendizado consiste no treinamento de dois níveis de aprendizado de máquina. Os aprendizes de nível 0 têm como função realizar previsões de valores faltantes no banco de dados. Estas previsões são entregues para o aprendiz de nível 1 (meta-aprendiz), que escolhe dentre os valores sugeridos, os mais adequados. Em nossa aplicação, o meta-aprendizado contém o acréscimo de técnicas de triangulação nos aprendizes de nível 0, visto que, usando dados de cidades vizinhas conseguimos encontrar dados faltantes com uma maior exatidão. Desta forma, no nível 0 é realizado o treinamento das máquinas usando os dados da cidade alvo e as técnicas de triangulação usando os dados de três cidades vizinhas. Os resultados dos dois processos são enviados aos meta-aprendizes (nível 1) que escolherão os valores mais adequados para a imputação.

Este artigo destina-se a apresentar uma ferramenta desenvolvida inicialmente por [Alves and Pereira 2020, Coelho et al. 2021] e aperfeiçoada pelos autores do trabalho atual.

2. Metodologia

2.1. Aquisição e Tratamento de Dados

Os dados utilizados na aplicação são obtidos junto ao site do INMET. Neste trabalho, são utilizados dados diários de temperatura máxima, temperatura mínima e precipitação total. Na maioria das estações encontramos dias com ausências de dados para um ou mais parâmetros. Um dos objetivos deste trabalho é tratar estas ausências de dados.

```
Nome: BELO HORIZONTE
Codigo Estacao: 83587
Latitude: -19.93444444
Longitude: -43.95222222
Altitude: 915.47
Situacao: Operante
Data Inicial: 1960-01-01
Data Final: 2022-03-17
Periodicidade da Medicao: Diaria

Data Medicao;EVAPORACAO DO PICHE, DIARIA(mm);INSOLACAO TOTAL, DIARIO(h);PRECIPITACAO TOTAL, DIARIO(mm);TEMPERATURA MAXIMA, DIARIA(°C);TEMPE
1961-01-01;1;1;30;3;24;8;21;18;6;93;25;0;
1961-01-02;1;4;10;6;23;8;20;32;18;8;97;5;0;
1961-01-03;5;9;36;1;24;7;20;9;18;7;97;1;133333;
1961-01-04;5;0;38;1;21;4;19;96;18;8;98;5;866667;
1961-01-05;3;0;67;1;22;4;20;04;18;4;95;25;0;
1961-01-06;5;1;7;2;25;6;20;56;18;4;89;75;1;5;
1961-01-07;1;1;3;1;17;9;27;3;21;98;18;6;88;5;0;
```

Figura 1. Cabeçalho de uma base de dados de uma estação do INMET

¹<https://docs.python.org/3/library/tkinter.html>

²<https://portal.inmet.gov.br/>

As bases de dados das estações possuem em seu cabeçalho as coordenadas geográficas e a altitude das mesmas (Figura 1). Tais informações são de suma importância para as técnicas de triangulação. Desta forma, o tratamento de dados conta com duas etapas, onde a primeira consiste em extrair das estações o nome, a localização e os dados climáticos. Já a segunda etapa, resume-se em coletar os dados diários das quatro estações em que a data entre elas são em comum. Isso é importante para que os algoritmos de triangulação possam ser utilizados para a realizar a inferência dos dados da cidade alvo. Nos casos onde houver poucos dados em comum, o usuário poderá refazer o processo escolhendo outras cidades.

2.2. Análise de Dados

Após a aplicação realizar o tratamento dos dados, uma visualização individual dos dados ficará disponível na aplicação. Esta visualização é importante para analisar se tais localidades escolhidas possuem uma janela considerável de dados ausentes. Vale ressaltar que, na seção de visualização de dados é possível escolher qual indicador climático e qual período dos dados o usuário deseja visualizar (Figura 2).

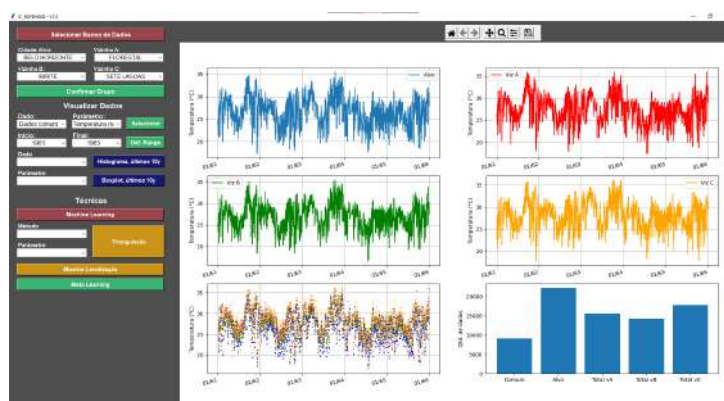


Figura 2. Exemplo de análise de dados para as cidades de BH, Florestal, Ibirité e Sete Lagoas, em MG

O sistema possibilita ainda outras duas formas de análise de dados, com o objetivo de observar a tendência climática. Este processo se inicia selecionando a localidade e o parâmetro climático, que tornará possível a opção de visualizar o histograma dos últimos 10 anos do banco de dados. Este tipo de disposição apresenta a distribuição de frequência dos valores, onde cada barra representa um número e a altura representa a frequência com que o valor ocorre. O sistema gera uma linha em vermelho, simbolizando a mediana do parâmetro selecionado para cada ano. Essa linha visa facilitar a visualização da tendência climática, visto que a mediana é uma medida de tendência central que indica o valor que divide o grupo de dados.

2.3. Aprendizado de Máquina

Na presente aplicação, encontra-se uma seção para modelagem e validação de aprendizado de máquinas, utilizadas tanto para imputação de dados pretéritos faltantes, quanto para fazer previsões de indicadores climáticos. São utilizados cinco algoritmos baseados na biblioteca ScikitLearn [Pedregosa et al. 2011], a saber: *Decision Trees*, *Nearest Neighbors*, *Support Vector*, *Gaussian Process* e *Neural Network*.

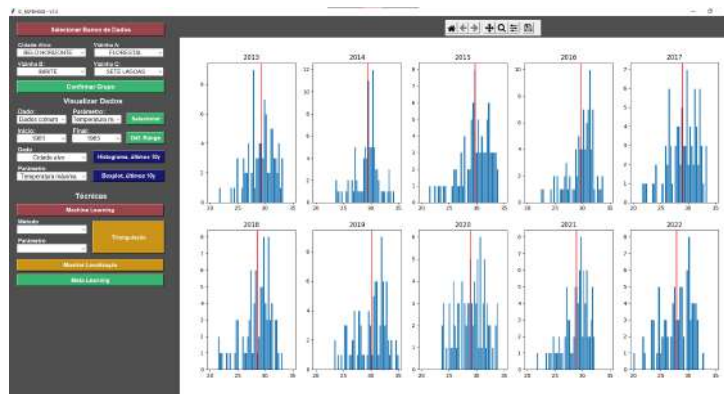


Figura 3. Histograma de temperatura máxima para Belo Horizonte-MG

Na parte de treinamento, foi utilizado o método de janela deslizante para realizar o *input* do treinamento e previsão em cada máquina. Tal método, consiste em percorrer pela série temporal e montando um subgrupo de dados, na qual nesta aplicação usamos um grupo de seis dias no total. Desta forma, estrutura dos dados de entrada consiste de: uma matriz (X), denominada de variável de entrada, onde cada linha consta seis datas e seus valores, porém, o valor do sexto dia é separado e armazenado em um vetor (Y), denominado de variável de saída. Vale salientar que, quando a máquina fizer a previsão de algum dia, é necessário entregar a data da previsão com 5 datas anteriores, juntamente com seus respectivos dados, como exemplo temos: $X = [2022/1/1, 20.5, \dots, 2022/1/5, 21.3, 2022/1/6]$, $Y = [22.8]$.

Para validação foi utilizado a abordagem *Hold-out*, que consiste em pegar a base de dados originais e dividir em duas porções, a de treinamento e de validação. Na seção de aprendizado de máquina, é sugerido a divisão de 70% para o treinamento e 30% para validação, mas lembrando que fica à escolha do usuário. Na seção de Aprendizado de Máquina (Figura 4), após o usuário escolher qual modelo irá utilizar, aparecerão os parâmetros do respectivo modelo. Quando o treinamento do modelo escolhido for finalizado, ficará disponível o resultado obtido pela máquina, contendo os erros absolutos, erros relativos e um gráfico que apresentará a comparação dos dados reais com os dados das previsões.

2.4. Triangulação

Outra abordagem que a ferramenta apresenta são as técnicas de triangulação, onde são usados dados e as coordenadas geográficas das cidades vizinhas para encontrar um dado ausente da cidade alvo. Nesta aplicação, usamos quatro métodos de triangulação discutidos no trabalho de [Alves and Pereira 2020], das quais são: *Arithmetic Average*, *Inverse Distance Weight*, *Optimized Normal Ratio* e *Regional Weighted*.

Podemos dizer que os resultados obtidos pelos métodos de triangulação são melhores em termos qualitativos e quantitativo (Figura 5), comparados com os resultados de aprendizado de máquina. Porém, este tipo de técnica fica preso à quantidade de dados das cidades vizinhas e a distância entre elas, enquanto as máquinas tem uma flexibilidade maior com apenas os dados da cidade alvo.

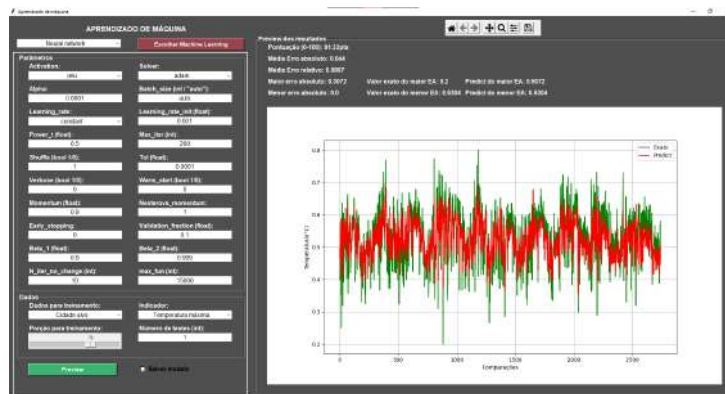


Figura 4. Tela de Aprendizado de Máquina

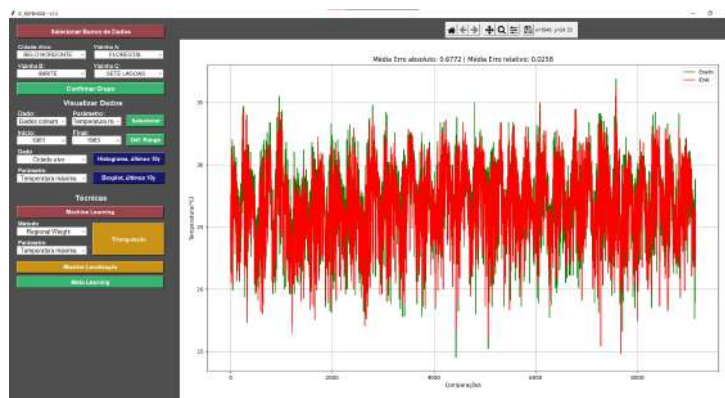


Figura 5. Resultado de Triangulação para a Cidade alvo usando o método Regional Weight

2.5. Meta Aprendizado

Outra técnica que este trabalho apresenta é o Meta Aprendizado, que consiste num método com dois níveis de modelos. Para este método, os dados foram divididos em porções em percentuais de 40%-40%-20%, para modelos de aprendizado de nível 0, de nível 1 e para a validação, respectivamente. No modelo de nível 0 temos os aprendizes básicos, aqueles usados em Machine Learning e também nos métodos de triangulação. Dessa maneira, 40% dos dados são entregues para estas funções com objetivo de realizar o treinamento e a outra porção para realizar previsões. Portanto, os resultados do processo dos aprendizes de base (nível 0) são entregues juntamente com a segunda porção de 40% para o aprendiz de nível 1, que correspondem às outras máquinas. Esses aprendizes de nível 1, utilizam os dados oriundos dos aprendizes de base para realizar seu treinamento. E, com a porção de 20% validamos os resultados.

A fim de realizar combinações entre aprendizes, a ferramenta disponibiliza um teste personalizado, onde é possível escolher os aprendizes dos dois níveis, e também conta com um teste global, onde o algoritmo realizará todas as combinações possíveis de aprendizes. Como trata-se de uma sequência de vários testes, a execução desta função pode demorar alguns minutos. Como resultado do teste global, temos uma lista com todas

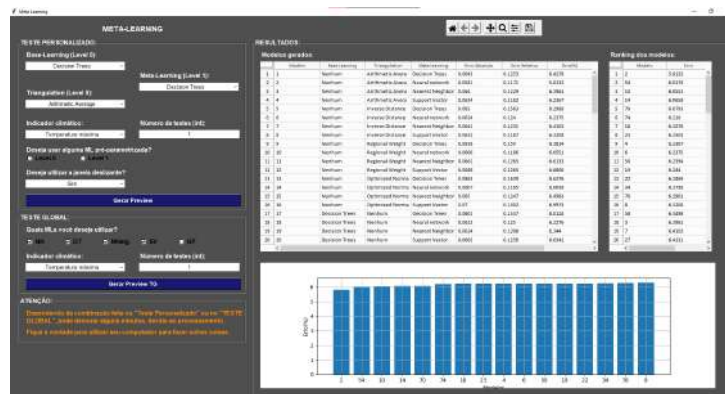


Figura 6. Resultados de um Teste Global no Meta Aprendizado

as possibilidades de modelos, garantindo a visualização de qual combinação apresentou os melhores resultados (Figura 6).

3. Conclusões

Diversas áreas da sociedade necessitam utilizar dados climáticos em suas atividades para se obter melhores resultados. O objetivo da aplicação apresentada neste trabalho é fornecer tanto para pesquisadores quanto para outros usuários, métodos de geração de dados meteorológicos de forma fácil e intuitiva.

Como nem todas as regiões possuem uma estação meteorológica, a aplicação pode produzir resultados de imputação para uma localização a partir de outras três estações, tornando-se uma alternativa nesse contexto, a fim de estimar dados para estas localidades específicas.

A aplicação encontra-se disponível para download no endereço https://github.com/PabloPassarini/Demo_toolbox_weather.

Agradecimentos

Os autores agradecem a PROPE/UFSJ e à FAPEMIG pelo suporte financeiro através do projeto APQ-00718-21.

Referências

- Alves, V. H. and Pereira, M. A. (2020). A meta-learning framework for imputing missing values in weather time series. pages 58–69. GeoInfo2020.
- Coelho, F. J. S., de A. Pereira, M., Davis-Jr., C. A., Silva, N. G., and Silva, T. T. D. (2021). Study on changing trends in climatic extremes in the brazilian territory. In *XXII Brazilian Symposium on Geoinformatics - GeoInfo 2020, São José dos Campos, SP, Brazil, online, November 29 - December 2, 2020*, pages 144–155. MCTIC/INPE.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

OpenStreetMap Tools for Humanitarian Aid

Alexander Zipf^{1,2}, Amandus Butzer², Marcel Reinmuth², Michael Auer², Benjamin Herfort², Adam Rousell², Sven Lautenbach^{1,2}, Stefan Gumbrich²

¹GIScience Research Group – Heidelberg University
INF 368 – 69120 Heidelberg – Germany

² HeiGIT gGmbH – Heidelberg Institute for Geoinformation Technology
Berliner Str. 45, 69118 Heidelberg.
zipf@uni-heidelberg.de

Abstract. *This paper describes a set of services based on OSM developed for humanitarian action at HeiGIT and GIScience Heidelberg. These are available at <https://github.com/GIScience>. In this tools paper we describe the latest developments for selected tools, like the OpenStreetMap history analysis framework (ohsome.org) and the mobility platform openrouteservice.org and their respective usage in new services like the ohsome quality analyst OQT, the global access healthcare map or a new HeiGIT disaster portal API.*

1. Introduction - the Potential and Challenge of OSM for humanitarian Aid

In addition to remote sensing data and other traditional sources for geographic information OpenStreetMap (OSM) has become a relevant complimentary dataset for humanitarian aid or disaster management (Herfort et al. 2021). Yet, the varying quality of the OSM data, especially in the Global South, still poses challenges for application and calls for specialized tools for analyzing, improving and better using OSM data for humanitarian and disaster management related purposes. Therefore HeiGIT and GIScience Heidelberg develop a growing set of Open Source tools, services and datasets related to those use cases applying OSM. We focus on the main two frameworks which are both important for supporting disaster management, humanitarian aid and similar use cases.

2. Towards reproducible Fitness-for-purpose OSM Quality Analytics in real time

The goal of ohsome is to make OpenStreetMap's full-history data more easily accessible for various kinds of data analytics tasks in an open and reproducible way on a global scale. Applications of the ohsome platform range from web dashboards over data quality assessment to custom data analysis. The ohsome framework can be used in a large range of OSM history and quality analytics, e.g. Auer et al 2018, Herfort et al 2021, Schott et al 2021, Grinberger et al 2021, Klonner et al 2021 etc. Recent additions and improvements concern several topics:

- the ohsome API <https://docs.ohsome.org/ohsome-api/v1/>
- and new R libraries <https://github.com/GIScience/ohsome-r>, as well as Python <https://github.com/GIScience/ohsome-py> and QGIS plugins <https://github.com/GIScience/ohsome-qgis-plugin> to use this API from your favorite computing environment,
- The new update mechanism for near realtime (minutely) updates of the underlying OSHDB (Raifer et al 2019) in the new Version 1.0 of OSHDB <https://github.com/GIScience/oshdb>

- new dashboard functionalities of both the ohsome dashboard <https://ohsome.org/apps/dashboard/> and the OSM history explorer ohsomeHEX <https://hex.ohsome.org/>
- the ohsome2label <https://github.com/GIScience/ohsome2label> (Wu et al 2020, Li & Zipf 2022), which can be used in applications such as to improve OSM missing building detection using deep learning as demonstrated for Africa in Li et al (2022a), detecting surface water (Li et al. 2021) or related infrastructure like wastewater treatment plants (Li et al. 2022b).
- and in particular the ongoing work of the ohsome quality analyst (OQT).
<https://oqt.ohsome.org/> OQT API Docs: <https://oqt.ohsome.org/api/docs>
 Sources: <https://github.com/GIScience/ohsome-quality-analyst>

The ohsome quality analyst (OQT) is a software to compute quality estimations on OpenStreetMap (OSM) data. End users interested in OSM quality can use the OQT to get hints on the quality of OSM data for their specific region and use case. OQT can be used via a web front-end as well from command line. Several indicators (quality measures) are combined in a report to describe the quality for one specific use case. We work on several new functionalities and indicators which will be described in the presentation. A recent addition includes indicators based on machine learning models, like the new building completeness indicator. OQT shall also benefit from the mentioned minutely updates of the OSHD as backend in the future, but also will rely on further more static auxiliary data sets, especially when those are integrated in advanced quality indicators, such as those based on machine learning models.

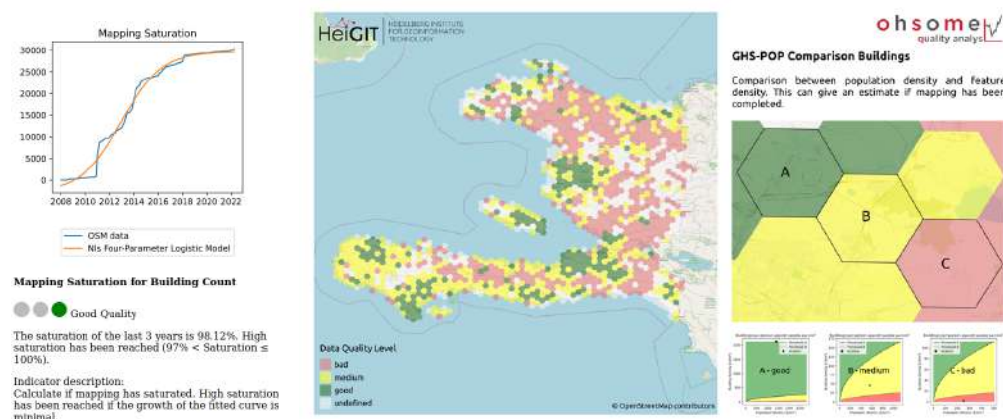


Figure 2. Examples of report results from the ohsome Quality Analyst (OQT)

3. Routing & Logistics for Disaster Management – towards automated workflows from satellite to logistics support

Openrouteservice (ORS) <https://openrouteservice.org/> has been used in disaster management for a long time (Neis et al 2008, 2010). More recent use cases include the investigation of the effect of flooding on the accessibility of healthcare facilities (Klipper et al. 2021), which combines ideas presented below like disaster routing and healthcare accessibility.

In general openrouteservice itself allows users to specify areas that should not be included in routing calculations (determination of a route from point A to point B, or in the calculation of reachability isochrones). Such areas could be things such as flooded areas or landslides, but in general identify areas that are not passable by the movement method indicated. For example disaster affected areas are displayed in the specific web client openrouteservice for disaster management. <https://disaster.openrouteservice.org/>.

If a user wants to exclude a certain area from the routing, they can interactively draw the corresponding area there as a so-called "AvoidArea" as a polygon. The high update frequency of the openrouteservice for disaster management also enables the consideration of very recent changes in OpenStreetMap (OSM). Here, too, changes, e.g. to the trafficability of roads, are quickly entered by the OSM community and can thus be used for routing. While this is a powerful tool, as openrouteservice itself does not provide any data storage capabilities, these polygons must be passed with each request. This can be problematic when multiple users want the same information to be used as they must manually share it between themselves.

A interesting new special service goes one step further. It takes the flooding data of the Copernicus EMS into account in the route and, on this basis, directly outputs a detour without the user having to provide any further information on those to be avoided. A first version has been used in the German floods 2021. We are currently developing this further and in particular are searching for further data providers of relevant information of non or less accessible areas due to disasters. This includes both local or regional organizations, but in particular National Mapping Agencies like INPE or disaster organizations like CEMADEN. Together we can develop workflows to achieve a more automated usage of disaster related geoinformation into disaster management applications related to logistics, routing, navigation or disaster-related planning.

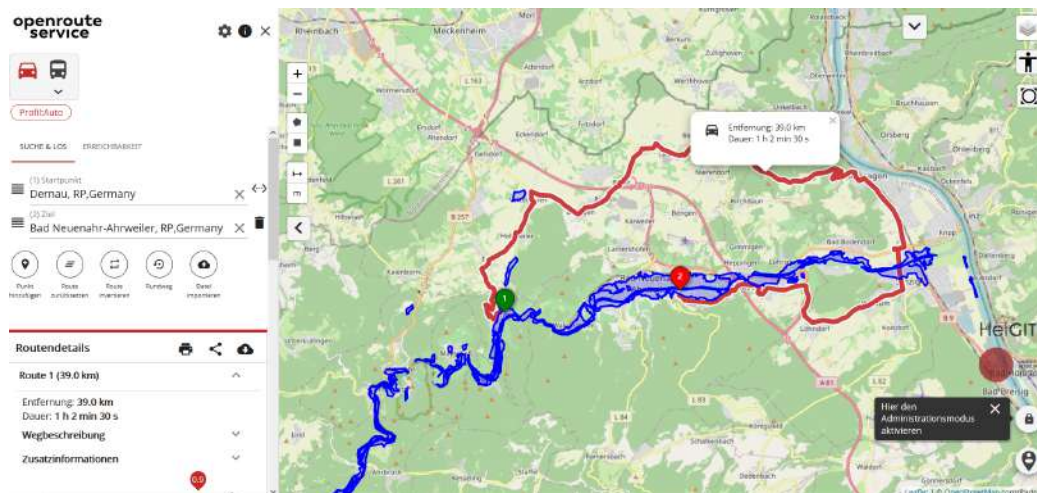


Figure 2. Example of ORS using Copernicus ERS flooded areas data as avoid areas via the HeiGIT disaster portal API at the floods in Germany

The current version of the application uses the new HeiGIT disaster portal API. This was developed to allow the accessing, adding, deleting and editing of geographic features that can be used in requests to openrouteservice or other services. Based on this initial development new data providers, workflows and functionalities can be added in future work. A first step may include specific data adapters for datasets provided by e.g. INPE or CEMDAEN to the current version of the HeiGIT disaster portal API.

But in the long run for example also the default speeds for routing could be adapted based on the severity of the damage. Currently we are targeting performance improvements for complex large vector data sets in order to improve real world applicability for large regions, given that many data providers publish raster data which needs to be transformed in suitable vector data sets beforehand in an automated way.

4. Global Open Healthcare Access Map

Recently we started the Open Healthcare Access Map. This web application combines data on healthcare infrastructure and motorized mobility to measure accessibility inspired by earlier work on Africa by Geldsetzer et al. (2020). Through the combination of accessibility (isochrones by openrouteservice (Neis & Zipf 2008) and population distribution, insights into health care supply are provided. The information is aggregated on administrative levels and hexagons. This differences in healthcare supply to be revealed spatially. When determining the accessibility of a location, a key consideration is not just how close it is, but also how long it takes to get there. In the case of driving, different road types and surfaces can affect travel speeds and this in turn affects how long it takes to get to the location. In the case of health care facilities, it is often the case that time is critical, and so having a realistic knowledge of travel times is essential. For example, in a number of medical conditions, the ability to reach a specialized facility within a particular time frame can be the difference between life and death. By knowing which areas are outside of the “reach” of these facilities, planners can better determine where focus needs to be placed on aspects such as the sites of new facilities, which facilities to optimize, or which road infrastructure should be improved.



Figure 1. Global Coverage of the Open Healthcare Access Map.

This app is intended as a prototype to showcase the potentials of free and open geodata within the public health domain. Future milestones are among others, the delineation of facility level catchment areas and incorporation of data quality measures. Initially, only a few countries and continents were featured. Now we are releasing a number of previously missing countries, achieving real global coverage.

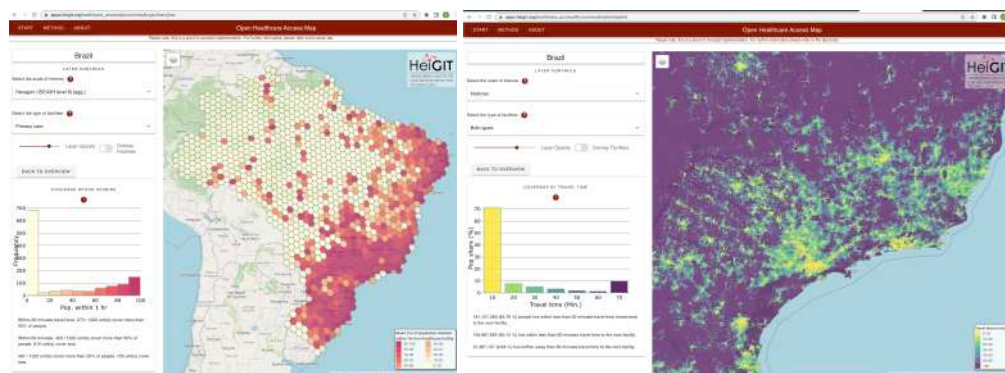


Figure 2. Access to healthcare at different scales for Brazil based on the global Open Healthcare Access Map. https://apps.heigit.org/healthcare_access/#/

A key objective of the project is to make the data more accessible to users. Therefore, we will make the data available on HDX. The first country we are publishing data for is Pakistan. Pakistan has been hit by severe flooding in 2022. The published datasets provide aggregated information at different scales on the coverage of the population with health services. The data can be used to compare the coverage of different regions and identify potential gaps in coverage.

5. Summary and Outlook

Further versions of all services shall be developed based on feedback and requirements from real world users, esp. researchers, NGOs, public authorities and stakeholders in humanitarian aid and disaster management. This includes both the assessment of the data quality of OpenStreetMap with the design of specific quality indicators and measures according to specific use cases (fitness for purpose, Barron et al 2013)), as well as the improvement of functionalities and development of tools for disaster logistics and routing. The quality of OSM needs always to be investigated for those use cases and in the future we will focus on quality aspects of the combination of routing related map features and attributes in particular. As demonstrated the the machine learning applications based on ohsome2label this also includes the generation of predictions for missing data or new specific data products for particular applications.

A further step goes into looking into new data sources and specific use cases: E.g. openrouteservice considers many characteristics of roads when determining travel speeds such as the classification (primary, secondary, street etc.) and the surface material (asphalt, cobblestone, bare earth etc.). In regions of the globe where OSM data has been heavily enriched with such information, travel times can be very accurate. However, it is often the case that in many regions such fine scale information is not available in the OSM dataset and so travel times are underestimated. Coupled with dynamic changes to driving conditions caused by heavy traffic or environmental influences (e.g. flooding, landslides), generating a realistic travel times and accessibility assessment is very difficult. We plan to tackle this problem together with partners around the world and try to investigate methods in order to find proxies for predicting better travel times. First results to build upon have been achieved in the project SocialMedia2Traffic (Zia et al. 2022). To adapt this to Brazil with local partners is of great interest to us. This can benefit also future cooperation with disaster management agencies and providers of disaster related data products such as INPE or CEMADEN.

6. References

- Auer, M.; Eckle, M.; Fendrich, S.; Griesbaum, L.; Kowatsch, F.; Marx, S.; Raifer, M.; Schott, M.; Troilo, R.; Zipf, A. (2018): Towards Using the Potential of OpenStreetMap History for Disaster Activation Monitoring. ISCRAM Conference 2018. Rochester. NY. US.
- Barron, C., Neis, P. & Zipf, A. (2013): A Comprehensive Framework for Intrinsic OpenStreetMap Quality Analysis. Transactions in GIS, DOI: 10.1111/tgis.12073.
- Geldsetzer, P.; Reinmuth, M.; O Ouma, P.; Lautenbach, S.; A Okiro, E.; Bärnighausen, T.; Zipf, A. (2020): Mapping physical access to health care for older adults in sub-Saharan Africa and implications for the COVID-19 response: a cross-sectional analysis The LANCET Healthy Longevity, Vol 1, Issue 1. [https://doi.org/10.1016/S2666-7568\(20\)30010-6](https://doi.org/10.1016/S2666-7568(20)30010-6)
- Neis, P. & Zipf, A (2008): OpenRouteService.org is three times “Open”: Combining OpenSource, OpenLS and OpenStreetMaps. (GISRUK 08). Manchester.
- Herfort, B., Lautenbach, S., Porto de Albuquerque, J., Anderson, J., Zipf, A. (2021): The evolution of humanitarian mapping within the OpenStreetMap

- p>community. Scientific Reports 11, 3037 (2021). DOI:10.1038/s41598-021-82404-z
- Raifer, M., Troilo, R., Kowatsch, F., Auer, M., Loos, L., Marx, S., Przybill, K., Fendrich, S., Mocnik, F.-B. & Zipf, A. (2019): OSHDB: a framework for spatio-temporal analysis of OpenStreetMap history data. *Open Geospatial Data, Software and Standards* 2019 4:3. <https://doi.org/10.1186/s40965-019-0061-3>
- Schott, M.; Grinberger, A. Y.; Lautenbach, S. & Zipf, A. (2021) The Impact of Community Happenings in OpenStreetMap – Establishing a Framework for Online Community Member Activity Analyses In: *ISPRS Int. J. Geo-Inf.* 10, no. 3: 164. <https://doi.org/10.3390/ijgi10030164>
- Grinberger, AY, Schott, M, Raifer, M, Zipf, A. (2021): An analysis of the spatial and temporal distribution of large-scale data production events in OpenStreetMap. *Transactions in GIS.* 2021; 00: 1– 20. <https://doi.org/10.1111/tgis.12746>
- Klipper, I.G., Zipf, A., Lautenbach, S., (2021): Flood Impact Assessment on Road Network and Healthcare Access at the example of Jakarta, Indonesia. *AGILE: GIScience Series* 2, 4, <https://doi.org/10.5194/agile-giss-2-4-2021>
- Klonner, C., Hartmann, M., Dischl, R., Djami, L., Anderson, L., Raifer, M., Lima-Silva, F., Degrossi, L. C., Zipf, A., de Albuquerque, J. P. (2021): The Sketch Map Tool Facilitates the Assessment of OpenStreetMap Data for Participatory Mapping. *ISPRS International Journal of Geo-Information.* 10:130. <https://doi.org/10.3390/ijgi10030130>
- Li, H. J. Zech, C. Ludwig, S. Fendrich, A. Shapiro, M. Schultz, A. Zipf (2021): Automatic mapping of national surface water with OpenStreetMap and Sentinel-2 MSI data using deep learning.. *International Journal of Applied Earth Observation and Geoinformation*, Vol 104, 2021, 102571. <https://doi.org/10.1016/j.jag.2021.102571>.
- Li, H. & Zipf, A. (2022): A conceptual model for converting openstreetmap contribution to geospatial machine learning training data, *Int. Arch. Photogramm. Remote Sens. Spatial Inf. Sci.*, XLIII-B4-2022, 253–259. DOI:10.5194/isprs-archives-XLIII-B4-2022-253-2022
- Li, H., Herfort, B., Lautenbach, S., Chen, J. & Zipf, A. (2022a): Improving OpenStreetMap missing building detection using few-shot transfer learning in sub-Saharan Africa. *Transactions in GIS*, 00, 1– 22. <https://doi.org/10.1111/tgis.12941>
- Li, H., Zech, J., Hong, D., Ghamisi, P., Schultz, M., Zipf, A. (2022b): Leveraging OpenStreetMap and Multimodal Remote Sensing Data with Joint Deep Learning for Wastewater Treatment Plants Detection. *International Journal of Applied Earth Observation and Geoinformation*, Volume 110, June 2022, 102804, <https://doi.org/10.1016/j.jag.2022.102804>
- Neis, P., Singler, P. & Zipf, A. (2010): Collaborative mapping and Emergency Routing for Disaster Logistics – Case studies from the Haiti earthquake and the UN portal for Afrika. In: Car, A., Griesebner, G. & Strobl, J. (Eds.): *Geospatial Crossroads @ GI_Forum '10. Proceedings of the Geoinformatics Forum Salzburg.*
- Wu, Zhaoyan, Li, Hao, & Zipf, Alexander. (2020). From Historical OpenStreetMap data to customized training samples for geospatial machine learning. In *proceedings of the Academic Track at the State of the Map 2020 Online Conference*, July 4-5 2020. DOI: <http://doi.org/10.5281/zenodo.3923040>
- Zia, M.; Fürle, J.; Ludwig, C.; Lautenbach, S.; Gumbrich, S.; Zipf, A. *SocialMedia2Traffic: Derivation of Traffic Information from Social Media Data.* *ISPRS Int. J. Geo-Inf.* 2022.

Index of authors

- Abner Anjos, 247
Adam Rousell, 441
Afonso Santos, 353, 429
Alber Sanchez, 192, 229
Alexander Zipf, 441
Alexandre Vidal, 111
Aline Soterroni, 211
Amandus Butzer, 441
Ana Candeias, 74
Ana Carolina Carneiro, 241
Ana Carolina Freitas, 311
Ana Carolina Pessôa, 1, 259
Ana Larissa Freitas, 123, 259, 372
Anderson Soares, 39
André Hirsch, 323, 366
André Martinon, 275
André Schardong, 395
Andréa Dos Santos, 323
Andréa Santos, 366
Andrés Velastegui-Montoya, 235, 253
Angelica Caseri, 335
Anielli Souza, 347
Antonio Miguel Monteiro, 360, 417
Antonio Paez, 63
Arisvaldo Junior, 395
Arthur Miller, 389
Aurelienne Jorge, 384
Aurora Yanai, 1
- Benjamin Herfort, 441
Bianca Bartolomei, 144
Bruno Batista, 144
Bruno Schultz, 281
- Carlos Brumatti, 329
Carlos Carvalho, 180
Carlos Felgueiras, 287
Carolina Pinho, 63
Celso Ribeiro, 204
Celso Silva-Junior, 259, 311
Claudia Santos, 287
- Clenya Lima, 341
Clodoveu Davis Jr, 168, 305
Cláudia Almeida, 347
Cláudio Campelo, 15
Cândida Leite, 372
- Daniel Braga, 204
Daniel Calbino, 366
Daniel Pinheiro, 323
Diego Matos, 192
Diego Melo, 293
Diego Tomasiello, 217
Dário Hossoda, 395
Débora Dutra, 1
- Elbert Macau, 317
Eliana Vale, 389
Elton Escobar-Silva, 347
Eurico Paula, 275
- Fabiana Zioti, 247
Fabiano Morelli, 264
Felipe Souza, 99, 270
Felipe Xaud, 341
Fernanda Ramalho, 168
Fernando Pimenta, 323, 366
Fernando Teixeira, 323, 366
Filipe Santos, 323, 366
Flaris Feller, 51
Flavia Feitosa, 63, 217, 241
Flavia Jacobsem, 395
Flávio Santos, 27
Francisco Pereira, 341
Francisco Rodrigues, 335
Fábio Moura, 27
- Gabriel Alves, 281
Gabriel Pereira, 323, 366
Gabriel Sansigolo, 247
Gabriel Soares, 395
Geancarlo Guerrero-Cabrera, 253
Gilberto Queiroz, 156, 247, 264, 293, 347

Giovanni Dolif, 389
Gisele Milare, 293, 417
Glauciana Ataíde, 323, 366
Guilherme Avansi, 111
Guilherme Mataveli, 229
Gustavo Marcatti, 323, 366

Haron Xaud, 341
Helio Pedrini, 111
Hugo Rivera-Torres, 235

Igor Broggio, 372
Isaac Miranda Jr, 353, 429
Ivair Gomes, 366
Ivanildo Barbosa, 299
Izabelly Costa, 384

Jaqueline Soares, 223
Jeferson Feitosa, 378
Jeferson Mendes, 411
José Júnior, 323, 366
José Simões, 270
João Guerrero, 229
João Reis, 123, 259, 311
Jugurta Lisboa-Filho, 329
Juliana Marquez, 253
Julio Reis, 329

Karine Ferreira, 156, 247, 264, 281

Laercio Namikawa, 133, 287
Leonardo Matos, 27
Leonardo Prado, 323, 366
Leonardo Santos, 223, 317, 360, 378, 384, 401, 406, 411
Leonardo Silvestre, 86
Leticia Cabral, 360
Letícia Bomfim, 111
Liana Anderson, 1, 123, 204, 259, 311, 347, 372
Lina Olivos, 395
Lucas Lamounier, 144
Luis Cunha, 63
Luis Felipe Cunha, 241
Luis Sadeck, 235
Luiz Aragão, 1, 204
Luiz Carvalho, 281
Lígia Barbosa, 353, 429
Livia Tomás, 378, 401, 406

Malcon Mora-Araus, 253
Marcel Reinmuth, 441
Marcela Ferreira, 323, 366

Marconi Cunha, 353, 429
Marconi Pereira, 86, 323, 366, 435
Marcos Quiles, 384
Marcos Silva, 27
Maria Barbosa, 123, 347
Maria Graziotto, 133
Maria Isabel Escada, 417
Mariana Schaefer, 329
Marie-Paule Bonnet, 156
Marielcio Lacerda, 293
Mario Pujatti, 86
Mateus Catrinque, 305
Mateus Macul, 264
Maxwell Oliveira, 15
Mayra Brasco, 180
Maíra Matias, 417
Melise Paula, 144
Michael Auer, 441
Michael Diniz, 223
Michel Chaves, 229
Ming Tang, 317
Márcia Dompieri, 27

Nathália Carvalho, 259
Natã Silva, 435

Pablo Passarini, 435
Paulo Graça, 1
Paulo Gurgel, 287
Paulo Leão, 259, 311
Pedro Andrade, 211
Pedro Silva, 323, 366
Philip Fearnside, 1

Rafael Frinhani, 144
Raquel Oliveira, 299
Renata Quevedo, 235
Renato Castro, 323, 366
Ricardo Silva, 1, 204
Roberta Magalhães, 217
Roberto Almeida, 144
Rogerio Flores Júnior, 156
Rogério Chagas, 395
Rogério Negri, 99, 270, 378, 411
Ronaldo Mello, 51

Sabrina Gomes, 74
Sabrina Marques, 211
Salatíel Silva, 15
Sergio Rosim, 287, 423
Sidnei Santanna, 287
Silvana Marcellini, 395

Stefan Gumbrich, 441
Stephan Stephany, 275
Sven Lautenbach, 441
Swanni Alvarado, 311
Sávio Oliveira, 39

Telles Silva, 435
Thais Medeiros, 123
Thales Körting, 293, 347
Thiago Barreto, 423
Thiago Rodrigues, 281
Tiago Carneiro, 192
Tiago Ferraz, 311

Vagner Rodrigues, 39
Vander Freitas, 317
Vanessa Souza, 144
Vinícius Gino, 99, 270
Vitor Gomes, 156
Vitor Hossaki, 378
Viviana Herrera-Matamoros, 235

Wanderson Santos, 259
Webert Pinto, 323, 366
Wellington Martins, 39
Wender Silva, 341
Wesley Campanharo, 264

Yadira Jaramillo-Lindao, 253
Yong Zou, 317