MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÕES

urlib.net/www/2022/02.01.16.47-TDI

COMPLEX NETWORK METRICS IN A METEOROLOGICAL CONTEXT

Aurelienne Aparecida Souza Jorge

Master's Dissertation of the Graduate Course in Applied Computing, guided by Drs. Leonardo Bacelar Lima Santos, and Izabelly Carvalho da Costa, approved in January 28, 2022.

URL of the original document:
<http://urlib.net/QABCDSTQQW/46A2GKP>

INPE
São José dos Campos
2022

PUBLISHED BY:

Instituto Nacional de Pesquisas Espaciais - INPE
Coordenação de Ensino, Pesquisa e Extensão (COEPE)
Divisão de Biblioteca (DIBIB)
CEP 12.227-010
São José dos Campos - SP - Brasil
Tel.:(012) 3208-6923/7348
E-mail: pubtc@inpe.br

MINISTÉRIO DA
CIÊNCIA, TECNOLOGIA
E INOVAÇÕES

urlib.net/www/2022/02.01.16.47-TDI

COMPLEX NETWORK METRICS IN A METEOROLOGICAL CONTEXT

Aurelienne Aparecida Souza Jorge

Master's Dissertation of the Graduate Course in Applied Computing, guided by Drs. Leonardo Bacelar Lima Santos, and Izabelly Carvalho da Costa, approved in January 28, 2022.

URL of the original document:
<http://urlib.net/QABCDSTQQW/46A2GKP>

INPE
São José dos Campos
2022

Cataloging in Publication Data

Jorge, Aurelienne Aparecida Souza.

J768c Complex network metrics in a meteorological context / Aurelienne Aparecida Souza Jorge. – São José dos Campos : INPE, 2022.

xxiv + 61 p. ; (urlib.net/www/2022/02.01.16.47-TDI)

Dissertation (Master in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2022.

Guiding : Drs. Leonardo Bacelar Lima Santos, and Izabelly Carvalho da Costa.

1. Networks. 2. Graph theory. 3. Meteorological radar. 4. Precipitation. I.Title.

CDU 004.7:004.652

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÕES

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

DEFESA FINAL DE DISSERTAÇÃO DE AURELIENNE APARECIDA SOUZA JORGE
BANCA Nº 007/2022, REG 138134/2019

No dia 28 de janeiro de 2022, as 09h, por teleconferência, o(a) aluno(a) mencionado(a) acima defendeu seu trabalho final (apresentação oral seguida de arguição) perante uma Banca Examinadora, cujos membros estão listados abaixo. O(A) aluno(a) foi APROVADO(A) pela Banca Examinadora, por unanimidade, em cumprimento ao requisito exigido para obtenção do Título de Mestre em Computação Aplicada. O trabalho precisa da incorporação das correções sugeridas pela Banca Examinadora e revisão final pelo(s) orientador(es).

Novo título: “Complex network metrics in a meteorological context”

Membros da banca:

Dr. Marcos Gonçalves Quiles - Presidente - Unifesp
Dr. Leonardo Bacelar Lima Santos - Orientador - Cemaden
Dra. Izabelly Carvalho da Costa - Orientadora - INPE
Dr. Rafael Duarte Coelho dos Santos - Membro Interno - INPE
Dr. Vander Luiz de Souza Freitas - Membro Externo - UFOP

Documento assinado eletronicamente por **Rafael Duarte Coelho dos Santos**, **Tecnologista**, em 01/02/2022, às 08:26 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Leonardo Bacelar Lima Santos**, **Pesquisador**, em 01/02/2022, às 08:51 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Marcos Gonçalves Quiles (E)**, **Usuário Externo**, em 01/02/2022, às 09:13 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Vander luis de souza freitas (E)**, **Usuário Externo**, em 01/02/2022, às 17:33 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

Documento assinado eletronicamente por **Izabelly Carvalho da Costa**, **Tecnologista**, em 11/02/2022, às 15:02 (horário oficial de Brasília), com fundamento no § 3º do art. 4º do Decreto nº 10.543, de 13 de novembro de 2020.

A autenticidade deste documento pode ser conferida no site http://sei.mctic.gov.br/verifica.html, informando o código verificador **9290078** e o código CRC **C5B5A627**.

Referência: Processo nº 01340.000523/2022-81 SEI nº 9290078

"The important thing is not to stop questioning."

ALBERT EINSTEIN

To my family, for being my basis

ACKNOWLEDGEMENTS

This work has become possible thanks to the support of several people I could count on throughout the entire trajectory.

First of all, I thank my parents, Marcos and Luciene, for the education they could provide me, always showing the great value of studies.

I am immensely grateful to my advisors, Dr. Leonardo and Dr. Izabelly, for all their support and guidance, always so solicitous. They are people with whom I learned a lot and motivated me in this initial trajectory in the scientific field.

I also register my thanks to all CAP professors for sharing their knowledge and making their time available.

And to every colleague that I could exchange experiences during this time, my gratitude for all the support and making the journey more pleasant.

I also express my gratitude to Dr. Vander and Iuri, from the Federal University of Ouro Preto, for all the valuable conversations and collaborations.

I thank my husband, Márcio, for his encouragement, patience, and being that partner I can always count on.

And to my son, Gael, who was born during this master's degree, my special thanks for giving me even more strength to conclude this research.

ABSTRACT

The study of Complex Networks represents an essential contribution to science as a tool to describe the structure of a wide range of complex systems in nature. Concerning atmospheric sciences, complex networks have been applied to climate data analysis, dealing with long-period and low-resolution data. Only a few works have been held in the weather domain, dealing with short-term changes in the atmosphere and manipulating spatial and temporal high-resolution data. What are the geographical and temporal signatures of meteorological processes in precipitation networks? To answer that, we present three case studies analyzing the behavior of network structures related to precipitation time series. The first one approaches the relations between topological and geographical distances and the spatial dependence inherent in a precipitation network. Our second case study compares different similarity measures and criteria for building up the networks, resulting in geographical findings: networks from distinct criteria occupying different spatial positions on a watershed, with a small number of shared edges. Finally, the last case study presents a description of the relations between topological metrics and meteorological properties in a series of precipitation events. As a result, we have explored the characteristics of meteorological networks in distinct scenarios, considering their spatial and temporal components. This way, we have prepared a basis for future research involving the application of complex networks to anticipate extreme weather events.

Keywords: Networks. Graph Theory. Meteorological Radar. Precipitation.

MÉTRICAS DE REDES COMPLEXAS EM UM CONTEXTO METEOROLÓGICO

RESUMO

O estudo das Redes Complexas representa uma contribuição importante à ciência como ferramenta para descrever a estrutura de uma variedade de sistemas complexos na natureza. No que se refere à área das Ciências Atmosféricas, as redes complexas têm sido aplicadas na análise de dados climáticos, envolvendo longas séries temporais e dados com baixa resolução. Até então, apenas algumas pesquisas foram realizadas na escala do tempo, tratando mudanças a curto prazo na atmosfera e manipulando dados com alta resolução espacial e temporal. Quais são as assinaturas geográficas e temporais de processos meteorológicos em redes de precipitação? Para responder a essa pergunta, apresentamos três estudos de caso analisando o comportamento de estruturas de rede relacionadas a séries temporais de precipitação. O primeiro aborda as relações entre as distâncias topológicas e geográficas e a dependência espacial inerente a uma rede de precipitação. No nosso segundo estudo de caso, comparamos diferentes métricas de similaridade e critérios para a construção das redes, com descobertas geográficas como resultado: redes de distintos critérios ocupando diferentes posições espaciais em uma bacia, com um pequeno número de arestas compartilhadas. Por fim, o nosso último estudo de caso apresenta uma descrição das relações entre métricas topológicas e propriedades meteorológicas em uma série de eventos de precipitação. Como resultado, foram exploradas as características de redes meteorológicas em distintos cenários, considerando as componentes espacial e temporal. Dessa forma, preparamos uma base para pesquisas futuras envolvendo a aplicação de redes complexas na antecipação de eventos de tempo extremo.

Palavras-chave: Redes Complexas. Teoria dos Grafos. Radar Meteorológico. Precipitação.

LIST OF FIGURES

Page

1.1 Global network connections. 1

4.1 Coverage of São Roque weather radar, the Metropolitan Area of São Paulo (yellow) and the Tamanduateí Basin (orange). 19
4.2 G4G Flowchart. 21

5.1 G4G flow for Case Study 1. 25
5.2 Temporal Correlation versus Geographical distance between each pair of points. Correlation values are grouped into three categories - minimum, medium and maximum values - respectively coloured in red, green and blue. 26
5.3 Topological distance versus Geographical (euclidean) distance. 27
5.4 Geographical network for Tamanduateí Basin. The white points represent the nodes, the blue segments are the edges of the network, and the yellow border is the outline of the basin. 27

6.1 G4G flow for Case Study 2 — pcGT and miGT networks. 31
6.2 G4G flow for Case Study 2 — pcBB and miBB networks. 32
6.3 G4G flow for Case Study 2 — pcCM and miCM networks. 33
6.4 Weight distribution for PC networks. 35
6.5 Weight distribution for MI networks. 36
6.6 Topological *versus* Geographical distance for each pair of nodes comparing pcGT *versus* pcBB — auxiliar lines representing linear regression, for pcGT (blue) and pcBB (green). 37
6.7 Topological *versus* Geographical distance for each pair of nodes comparing miGT *versus* miBB — auxiliar lines representing linear regression, for miGT (blue) and miBB (green). 38
6.8 Networks on the watershed (contour in yellow): pcGT (blue), pcBB (green), and shared edges (red). In the background, SRTM altimetric data (the lower a cell, the darker it is). 39
6.9 Networks on the watershed (contour in yellow): miGT (blue), miBB (green), and shared edges (red). In the background, SRTM altimetric data (the lower a cell, the darker it is). 40

7.1 MASP and the delimitation of the study area. 46
7.2 G4G flow for Case Study 3. 47

7.3 Flow to identify correlations between network and meteorological metrics. 47
7.4 Meteo-Network Graph: Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively). 49
7.5 Groups of events and their intersections. 50
7.6 Meteo-Network Graph - Group D1 (Short Duration): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively). 51
7.7 Meteo-Network Graph - Group D2 (Long Duration): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively). 52
7.8 Meteo-Network Graph - Group A1 (Short Extension): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively). 53
7.9 Meteo-Network Graph - Group A2 (Long Extension): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively). 54

LIST OF TABLES

Page

3.1 Comparison of the datasets mentioned in the literature review. 15
3.2 Tools for network creation and analysis. 17

6.1 Topological properties of each network. 34

7.1 Groups of Events. 50

LIST OF ABBREVIATIONS

BB	–	Backbone
CAPPI	–	Constant Altitude Plan Position Indicator
CPTEC	–	Centro de Previsão de Tempo e Estudos Climáticos
DECEA	–	Departamento de Controle do Espaço Aéreo
GT	–	Global-Threshold
MI	–	Mutual Information
NCEP	–	National Centers for Environmental Prediction
PC	–	Pearson Correlation
PPI	–	Plan Position Indicator
MASP	–	Metropolitan Area of São Paulo
Tathu	–	Tracking and Analysis of Thunderstorms

LIST OF SYMBOLS

N	–	number of nodes in a graph
L	–	number of edges in a graph
k_i	–	degree of node i
<k>	–	average degree
c_i	–	clustering coefficient of node i
<c>	–	average of the clustering coefficients
$<l_i>$	–	average of the shortest paths from node i
<l>	–	average shortest path of a graph
D	–	graph diameter
κ	–	heterogeneity parameter
NC	–	number of components
GC	–	giant component
ST	–	singletons

CONTENTS

Page

1 INTRODUCTION 1
1.1 Objective 2
1.2 Dissertation structure 3

2 THEORETICAL FOUNDATIONS 5
2.1 Climate and weather 5
2.2 Graph theory and network science 6
2.3 Similarity functions 7
2.4 Building up networks 8
2.5 Network metrics 9
2.6 Small world effect 11

3 LITERATURE REVIEW 13
3.1 Complex networks in climate and weather 13
3.2 Tools for network creation and analysis 16

4 METHODOLOGY 19
4.1 Data 19
4.2 (geo)graphs: geographical graphs 20
4.3 From binary data into geographical networks 20

5 FIRST CASE STUDY: SPATIAL DEPENDENCE AND RELATIONS BETWEEN TOPOLOGICAL AND GEOGRAPHICAL DISTANCES 23
5.1 Introduction 23
5.2 Materials and methods 24
5.2.1 Data 24
5.2.2 Network construction and analysis 24
5.3 Results 25
5.4 Final considerations 28

6 SECOND CASE STUDY: BUILDING UP DIFFERENT NETWORKS IN A WATERSHED 29
6.1 Introduction 29

6.2 Material and methods 30
6.2.1 Weather radar dataset and the Tamanduateí basin 30
6.2.2 The construction and analysis of the geographical networks 30
6.3 Results and discussion 34
6.4 Final considerations 40

7 THIRD CASE STUDY: ANALYSIS OF PRECIPITATION EVENTS AND RELATIONS BETWEEN NETWORK METRICS AND METEOROLOGICAL PROPERTIES 43
7.1 Introduction 43
7.2 Material and methods 44
7.2.1 Data 44
7.2.2 Study area 45
7.2.3 Precipitation event networks 45
7.3 Results and discussion 48
7.4 Final considerations 53

8 FINAL REMARKS 55

REFERENCES 57

1 INTRODUCTION

The study of Complex Networks represents an essential contribution to science to describe the structure of complex systems in nature. In such a description, the nodes represent the systems' components, and the edges describe the interactions between its components. Among its many applications, a commonly cited example is the mapping of the Internet network, consisting of routers, computers, and many other devices (nodes) connected by physical links (edges). Hydrographic networks, road systems, and social networks are other examples of applications in complex networks (WATTS; STROGATZ, 1998; STROGATZ, 2001; ALBERT; BARABÁSI, 2002; BARABÁSI; PÓSFAI, 2016).

Figure 1.1 - Global network connections.

SOURCE: Linforth (2020).

In the scope of nature, more specifically in the atmospheric sciences, the study of complex networks has been applied to the analysis of climatic data to identify structural patterns and teleconnections. To this end, researchers use long series of atmospheric variables from different sources, ranging from months to several years. They arrange such data in network structures based on some similarity measure: correlation, event synchronization, or mutual information. (TSONIS et al., 2006; DONGES et al., 2009; STEINHAEUSER et al., 2010; PALUŠ et al., 2011; MALIK et al., 2012; BOERS et al., 2014; JHA; SIVAKUMAR, 2017; BOERS et al., 2019).

Some papers analyze, for example, several years of reanalysis data from the United

States National Centers for Environmental Prediction (NCEP). Addressing a global spatial domain, they find patterns of teleconnections or a structure of communities with an intrinsic climatological interpretation. (TSONIS et al., 2006; STEINHAEUSER et al., 2010).

Other works base their researches on long time series of rainfall estimates from satellite data. With the use of event synchronization measures in such time series, they identify the relationship between the occurrence of extreme precipitation events between distant areas within a given time interval (BOERS et al., 2014; BOERS et al., 2019).

However, while climate refers to the average of atmospheric conditions over a long period, the weather domain deals with the most immediate state of the atmosphere. As a result of extreme weather conditions, floods, tornadoes, thunderstorms, and hail precipitation can emerge, causing severe impacts to society, with social and economic losses. The monitoring of weather conditions, analyzing severe weather parameters in advance could mitigate the effects of such events (ENORÉ et al., 2018).

A few works have applied complex networks specifically to study weather events. Ceron et al. (2019) have published one of them. Their paper analyzes precipitation estimate data from a meteorological radar with high spatial and temporal resolution. The authors show significant results in the detection of communities, based on a short-term time series, with just ten days (CERON et al., 2019).

What remains unknown in the context of complex meteorological networks is the spatial and temporal behavior of topological indices in networks of meteorological phenomena. The present work seeks precisely to try to answer these questions, specifically concerning precipitation events.

1.1 Objective

The purpose of this research is to contextualize network indices, specifically in the meteorological scope, addressing short-term atmospheric change fields, both spatially and temporally.

In this context, the scientific question to be answered is: “What is the behavior of meteorological processes in precipitation networks?”. In response, we expect to find patterns in topological indices, considering the geographic and temporal components of the network. The final objective is to present a descriptive approach to the different features found.

There are perspectives to use the concepts obtained in this research as a preliminary basis for the future development of tools that could help the nowcasting process realized by the meteorologists from the National Institute for Space Research (INPE). Among the perspectives, we can mention the attempt to promote the anticipated classification of events, applying network analysis to forecast data. Forecast products based on radar images and data from numerical weather prediction models are some examples of data that can serve as a basis for this analysis.

1.2 Dissertation structure

The structure of this research work is described below.

- Chapter 2. Theoretical Foundations: Basic concepts of climatology and meteorology; Definition of the similarity functions applied in this work; Description of the adopted methods to build up networks; Definition of network metrics; Concepts of the small world phenomenon;
- Chapter 3. Literature Review: Complex networks applied in climate and weather scales; Tools for network creation and analysis;
- Chapter 4. Methodology: Information about the dataset; Concepts about (geo)graphs; Methodology applied in the construction of the network;
- Chapter 5. Case Study 1: Spatial dependence and relations between topological and geographical distances;
- Chapter 6. Case Study 2: Building up different networks in a watershed;
- Chapter 7. Case Study 3: Analysis of precipitation events and relations between network metrics and meteorological properties;
- Final Remarks.

2 THEORETICAL FOUNDATIONS

2.1 Climate and weather

The *weather* term refers to the most immediate state of the atmosphere, comprising its short-term variations (minutes to days) (AMERICAN METEOROLOGY SOCIETY (AMS), 2018). Cavalcanti et al. (2009) point out that the atmosphere, however, is highly complex, defying the most straightforward definitions. Various phenomena, such as cyclones, anticyclones, atmospheric waves, and cold fronts, are associated with the weather we experience in our daily lives, such as rain, heat, and cold.

Climate is usually defined as the average of time, or more strictly, as the statistical description of the average and variability of quantities over a long period. This period may range from months to thousands or millions of years. A classic period adopted when dealing with climate is 30 years, as defined by the WMO. Quantities are often surface variables such as temperature, precipitation, and wind (WORLD METEOROLOGICAL ORGANIZATION (WMO), 2019).

Precipitation, the focus of this work, is the condensation product of water vapor in the atmosphere, which falls from clouds due to the gravity effect. For the formation of clouds to occur, a lifting process must take place. The water vapor present in the air, when reaching higher altitudes, condenses and forms droplets that are converted into rainfall. Convergence, convection, topography, and cold fronts are examples of lifting mechanisms (CAVALCANTI et al., 2009).

Clouds can be classified according to their appearance, shape, and altitude, the two primary categories being: (NUGENT et al., 2019; NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION (NOAA), 2015):

- Cumuliform: Also known as convective clouds, they develop due to vertical movements from instability in the atmosphere. They can result in storms with heavy rain and electrical discharges.
- Stratiform: Clouds in horizontal, uniform, flat layers, which tend to spread over larger areas. They are typically formed when a layer of air reaches saturation but is thermodynamically stable. They can also form when a convective cloud encounters a stable layer and spreads out in a layered format. May result in light precipitation or drizzle.

Both can produce significant rainfall accumulations. The cumuliform clouds are usu-

ally associated with high precipitation values in a short period. Moreover, the rainfall from stratiform clouds may reach substantial accumulations in a more extended period due to its persistence.

From a climatological point of view, the rainfall regime in a given area is the dominant factor in defining the local climate. However, these rainfall events result from a series of events in very different temporal and spatial scales. This way, the events in a certain location can affect the weather in other regions. These distance interactions occur through a mechanism called teleconnections (CAVALCANTI et al., 2009). As mentioned in the previous chapter, complex networks can be applied in atmospheric data to identify and analyze teleconnections.

When analyzing precipitation on the meteorological scale, events such as floods and inundations can be monitored or even predicted in advance (ANDERSEN; SHEPHERD, 2013). Other environmental and social impacts can be avoided or mitigated through monitoring and forecasting intense rainfall.

2.2 Graph theory and network science

Network science has its mathematical basis in graph theory, whose origin comes from the 18th century in Könisberg, the capital of Eastern Prussia. Its peculiar arrangement of bridges - seven in total - gave birth to a puzzle: Can one walk across all seven bridges and never cross the same one twice? The problem was solved by Leonardo Euler, a swiss mathematician, in 1735 with a representation based on nodes and edges. He represented each of the four land areas with letters (nodes) and the bridges with lines (edges). Then, Euler observed that, if there was a path crossing all bridges, but never the same bridge twice, nodes with an odd number of edges should be the starting or the end point of such a path. A path that goes through all bridges must have only one starting and one end point. Therefore, such a path cannot exist on the Könisberg graph, which has four nodes with an odd number of edges (BARABÁSI; PÓSFAI, 2016).

Although Euler is considered to be the creator of graph-theoretical ideas, the term *graph* was actually mentioned for the first time in 1878 by James Joseph Sylvester (SYLVESTER, 1877-8). In a formal definition, a graph G is defined by a set V(G) of elements called *vertices*, a set E(G) of elements called *edges*, and a relation of incidence, which associates each edge with one or two nodes called its *ends*. An edge can be called a *loop* if its starting and end points are exactly the same vertex. The terms *nodes* and *links* are also used instead of *vertices* and *edges*, respectively

(TUTTE, 2001).

To understand a complex system, regardless of its nature, we need to know how its components interact with each other. In network science, we can do that by representing the systems' components and their interactions through nodes and links of a graph. Two basic network parameters are the number of nodes (N) and the number of links (L). The links of a network can be *directed* or *undirected*. If the interaction between two nodes occurs strictly in one direction, the link is directed. Otherwise, if the interaction happens no matter the direction, the link is undirected (BARABÁSI; PÓSFAI, 2016). In this research, we use only undirected links, because we do not distinguish the direction of the interactions.

A complete description of a network requires the identification of all its connections. The simplest way to achieve that is through an adjacency list, whose elements are the pairs of nodes connected by the network's edges. There is also the option to represent the network through an adjacency matrix. Considering a directed network of N nodes, its adjacency matrix has N columns and N rows, its elements being:

$$A_{i,j} = \begin{cases} 1, & \text{if there is a link from node } i \text{ to node } j \\ 0, & \text{if nodes } i \text{ and } j \text{ are not connected to each other} \end{cases} \tag{2.1}$$

These element values are applied in networks whose all edges have the same weight. But in this work, we use weighted networks, which are networks where each edge has a unique weight w_{ij} ($A_{ij} = w_{ij}$).

2.3 Similarity functions

This section describes the similarity measures we use in this work, aiming to compare pairs of time series. The resulting values are the weights of our network edges. We employ two similarity measures: Pearson Correlation (PC) and Mutual Information (MI). The first is a normalized covariance between two series, capturing their joint variability. For a high PC coefficient, the variations in the values of one series must also happen in the other. Therefore, linear relations result in high coefficients. Its value is given by the Equation (2.2).

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{2.2}$$

Differently, the MI gives a notion of the shared information between the two series. It measures how much knowing one of them reduces the uncertainty related to the other. One could define MI in terms of conditional and joint entropy or joint and marginal probability functions (Equation (2.3)). Consequently, MI recognizes nonlinear relations as well (SHANNON, 1948) (KRASKOV et al., 2004).

$$I(X;Y) = \sum_{y \epsilon Y} \sum_{x \epsilon X} p(x,y) log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{2.3}$$

2.4 Building up networks

After applying a similarity measure to compare time series, an adjacency matrix is produced as an outcome. Based on that, we need to define a criterion to select the edges that will remain when building up the network. In this work, we use three different criteria for that purpose: Global Threshold (GT), Backbone (BB), and Configuration Model (CM).

In the GT criterion, the edges remain for those pairs whose similarity exceeds a global threshold, selecting only the highest similarity values. For this research, we define the global threshold value as the point of maximum diameter of the network. This way, we intend to promote the best possible balance between removing the least relevant edges and keeping the most important ones - as applied in previous papers in the literature (SANTOS et al., 2019; CERON et al., 2019). The probability that an edge of weight w_{ij} remains, given a global threshold value (gt), is:

$$p_{ij} = \mathcal{H}(w_{ij}) = \begin{Bmatrix} 1, w_{ij} \geq gt \\ 0, w_{ij} < gt \end{Bmatrix} \tag{2.4}$$

Distinctly, the BB approach uses each node's weight fluctuations to select the remaining edges. Considering a node i with degree k_i and strength s_i, we hypothesize that weights are randomly distributed over the k_i edges and sum up to s_i. The probability that an incident edge has weight w_{ij} or larger is (MENCZER et al., 2020):

$$p_{ij} = \left(1 - \frac{w_{ij}}{s_i}\right)^{k_i - 1}. \tag{2.5}$$

The edge is preserved if $p_{ij} < \alpha$. Since each edge has two endpoints, we consider the smaller p_{ij}. This approach selects edges that are incident to highly connected nodes

or those that substantially contribute to the node strength. The bigger the w_{ij} and k_i are, the smaller the resulting p_{ij} is.

The third criteria, CM, is based on a method available at the Networkx Library. It generates random pseudographs following a pre-defined degree distribution (NEWMAN, 2003). The main idea of using this approach is to have a null model for comparison.

2.5 Network metrics

There are some properties we can calculate related to nodes, edges, and even to the graph as a whole. Such properties help to explain the characteristics of a network structure. Next, we describe the metrics mentioned in this research. The descriptions and equations we present here refer specifically to undirected networks, which are the focus of this work.

The local metrics we employ, related to nodes individually or pairs of nodes, are:

- **Degree (k_i)**: number of connections of a given node; in an unweighted graph, the degree can be calculated based on the adjacency matrix:

$$k_i = \sum_{j=1}^{N} A_{ij} \tag{2.6}$$

- **Clustering Coefficient (c_i)**: indicates how connected a node's neighbors are to each other. For a node i with degree k_i, the local clustering coefficient is defined as ((BARABÁSI; PÓSFAI, 2016)):

$$c_i = \frac{2L_i}{k_i(k_i - 1)} \tag{2.7}$$

  Where L_i is the number of links between the neighbours of node i.

- **Shortest path length (l_{ij})**: the path with the minimum number of edges that connects the nodes i and j;

- **Node's average shortest path ($\langle l \rangle_i$)**: the average of the shortest paths from node i to all the other nodes (which are inside the connected compo-

nent):

$$\langle l \rangle_i = \frac{1}{N-1} \sum_{j=1, j \neq i}^{N} l_{ij} \tag{2.8}$$

And the following metrics are the ones we apply when analyzing the network as a whole:

- **Average degree ($\langle k \rangle$)**: average number of the degree metric of all nodes in the network:

$$\langle k \rangle = \frac{1}{N} \sum_{i=1}^{N} k_i \tag{2.9}$$

- **Network's average shortest path ($\langle l \rangle$)**: the average of all nodes' average shortest paths:

$$\langle l \rangle = \frac{1}{N} \sum_{i=1}^{N} \langle l \rangle_i \tag{2.10}$$

- **Average of the local clustering coefficients ($\langle c \rangle$)**: average of the clustering coefficients of all nodes in the network:

$$\langle c \rangle = \frac{1}{N} \sum_{i=1}^{N} c_i \tag{2.11}$$

- **Diameter (D)**: the longest shortest path of a network;

- **Heterogeneity parameter (κ)**: the ratio between the average of the squared degree and the square of the average degree:

$$\kappa = \langle k^2 \rangle / \langle k \rangle^2 \tag{2.12}$$

- **Number of components (NC)**: the number of isolated groups of nodes (components);

- **Giant component (GC)**: the size of the largest component;

- **Singletons (ST)**: the number of components with a single node.

2.6 Small world effect

In 1967, Stanley Milgram designed the first experiment to measure distances in social networks. He chose two individuals as targets. Then Milgram sent letters to randomly chosen people, asking them to forward these letters until they reached the target individuals. At the end of the experiment, he found the average number of 5.2 intermediate people before getting the letter to the final target. It is known as the *small world phenomenon* or *six degrees of separation* (ALBERT; BARABÁSI, 2002; BARABÁSI; PÓSFAI, 2016).

In network science, it implies that there is a short distance between two randomly chosen nodes in a network. It is possible to identify if a network structure could be statistically classified as a small-world network. We can do that by comparing the $\langle c \rangle$ and $\langle l \rangle$ between the original network and an equivalent random one - the same number of nodes and edges, but not necessarily preserving the degree distribution. Considering a graph with L edges randomly distributed by its N nodes, one can define analytically (ALBERT; BARABÁSI, 2002; BARABÁSI; PÓSFAI, 2016):

$$\begin{aligned} \langle k \rangle &= p(N-1), \\ \langle c \rangle &= p \quad and \\ \langle l \rangle &= log(N)/log(\langle k \rangle), \\ where \quad p &= 2L/[N(N-1)] \end{aligned} \tag{2.13}$$

If the original network has a lower $\langle l \rangle$ and a higher $\langle c \rangle$ than its equivalent random structure, it is possible to define it as a small-world network.

3 LITERATURE REVIEW

3.1 Complex networks in climate and weather

One of the first examples of the application of complex networks in climate, the article "What do networks have to do with climate?", Tsonis et al. (2006) reviews the literature on complex networks, presenting the network types and some of the metrics used to describe their structural properties. In order to extend these concepts to a climate system, the authors assume that a network could represent dynamic systems with interactions.

They use an NCEP reanalysis dataset with geopotential height values at a pressure level of 500 hPa to represent the atmosphere's general circulation. The data are monthly and cover the period from 1950 to 2004, with a spatial resolution of 5 degrees and a global domain (TSONIS et al., 2006).

When building up the network, each node represents a grid point, and it is related to a time series of anomaly values (calculated by subtracting the climatological average for each month). Next, they calculate the correlation coefficient between the time series. The results equal to or above 0.5 are considered to create the edges between the respective nodes (TSONIS et al., 2006). Once having constructed graphs, they measure some properties such as the clustering coefficient, which measures how connected the neighbors of a given node are, and the diameter, which represents the maximum shortest path between any pair of vertices in the network. Based on these metrics and comparing with equivalent random networks (networks with the same number of vertices and edges but with a random distribution of connections), the authors managed to classify the network according to its type. Furthermore, they revealed intrinsic characteristics of this climate network: stability and efficiency in transferring information (TSONIS et al., 2006).

Donges et al. (2009) constructed weather networks from surface air temperature data from reanalysis datasets and the coupled atmosphere-ocean general circulation model. Using different similarity metrics, they compared different spatial scales, highlighting the results using mutual information to detect network structures based on nonlinear physical processes.

Steinhaeuser et al. (2010) also apply the concepts of complex networks to climate data - NCEP reanalysis data over 60 years with monthly temporal resolution and 2.5º spatial resolution for a global domain - in order to identify regions with sim-

ilar climatologies. They use cross-correlation between atmospheric variables as the representative value of each grid point (node). Then, they apply the Euclidean distance between the correlation values of each node to map the similarities (weight of the edges). Their work employs community detection techniques to provide a climatological interpretation of the network structure.

Another example of a climate network is analyzed in Boers et al. (2014), which presents a framework for forecasting extreme precipitation events, focusing on the central-eastern Andes. They build up networks based on a nonlinear synchronization measure. As a result, they identify the spread of events from South America to the east-central region of the Andes within one day. Their dataset consists of 13 years of satellite data with a spatial resolution of 0.25º.

The nonlinear event synchronization measure is also applied to precipitation data from satellite images by Boers et al. (2019), using a dataset in a global spatial domain, from 1998 to 2016 with a daily temporal resolution. Analyzing the distances between the synchronized events, they identify the relationship between extreme precipitation events in the monsoon systems of south-central and eastern Asia and Africa.

Jha and Sivakumar (2017) also analyze precipitation data, but from a network of rain gauges, specifically from the Murray-Darling basin region in Australia. Data comprise the period from 1951 to 2014 with a daily temporal resolution. For this study, such data are grouped in different temporal scales to evaluate the impact of this variation on the structure of networks. The correlation threshold employed to select the edges of each network is also varied to evaluate the results.

Akbar and Saritha (2021) present a quantum inspired community detection that was successful in demonstrating an association between biodiversity change, climate change and land-use conversion. To analyze the climate change, the authors employ annual temperature and precipitation anomalies from 2010, 2014 and 2018.

Agarwal et al. (2022) investigate the pattern of extreme precipitation events in a river basin located in India. To do that, they employ two event-based nonlinear similarity measures: event synchronization and edit distance. The dataset includes a gridded daily precipitation data with a spatial resolution of 0.25°, which was produced from an interpolation of 6995 gauging stations. The time series contain 22 years (1998 to 2019) in a daily resolution.

Table 3.1 - Comparison of the datasets mentioned in the literature review.

Paper	**Variable**	**Source**	**Temporal Resolution and Domain**	**Spatial Resolution and Domain**
Tsonis et al. (2006)	Geopotencial Height at 500 hPa	NCEP reanalysis dataset	1950-2004 Monthly	Global 5º
Steinhaeuser et al. (2010)	Air Temperature, Pressure, Relative Humidity, Precipitable Water	NCEP reanalysis dataset	1948-2007 Monthly	Global 2,5 º
Boers et al. (2014)	Precipitation	Satellite (TRMM 3B42V7) + Rain Gauges	2001- 2013 3 hours	South America 0,25º
Jha and Sivakumar (2017)	Precipitation	Pluviometers	1951-2014 Daily	Murray-Darling Basin (Australia)
Boers et al. (2019)	Precipitation	Satellite (TRMM 3B42V7)	1998-2016 Daily	Global 0,25º
Ceron et al. (2019)	Precipitation	Weather Radar at Pico do Couto (DECEA)	24/01/2012 - 02/02/2012 10 minutes	Mountaineous Region of RJ 0,009º
Agarwal et al. (2022)	Precipitation	Gauging Stations	1998-2019 Daily	Ganga River Basin (India) 0,25º

All cited works apply complex network analysis to atmospheric data, dealing specifically with a climatic scale. Few works use complex networks in the weather domain, approaching a dataset that refers to the most immediate state of the atmosphere and, therefore, analyzes short-term changes in space and time. The article by Ceron et al. (2019) follows this line precisely and analyzes meteorological networks with high spatial and temporal resolution. The authors use community detection techniques, and as a result, present well-defined structures consistent with the region's topography and land use/cover. However, the authors use a single time series com-

prising ten days, not analyzing whether the network presents any dynamic behavior, which could allow, for example, classifications or even predictions.

Table 3.1 presents the characteristics of the data used by the mentioned works, making a comparison between the domains and spatial and temporal resolutions adopted.

3.2 Tools for network creation and analysis

There are several tools available whose goal is to construct and analyze networks. The options include programming language packages, libraries, database management systems, and even independent platforms. This section describes a few of the most commonly used tools.

Igraph is a well-known free and open-source package, often used in complex networks scenario. It is available in Python, C, R, and Mathematica languages. The package offers diverse functions for analyzing networks, offering good performance, portability, and ease of use (IGRAPH, 2020).

Networkx is a library created explicitly for use in Python language, which allows the creation, manipulation, and study of the dynamics and structure of complex networks. It allows the network construction based on different data types, such as text, image, and XML (NETWORKX, 2019).

Tidygraph is an R package that provides an API for graph manipulation. It encapsulates a good part of igraph functionalities, also adding some methods for relational data manipulation, such as graph joining. Internally, it structures the network by organizing nodes and edges in virtual data frames and offers an API for manipulating this data (PEDERSEN, 2019).

Stplanr, dodgr, and *spnetwork* are other R language packages that provide network analytical tools. Among them, the dodgr package is more specific to road networks, with a focus on directed and weighted graphs (LOVELACE; ELLISON, 2018; PADGHAM, 2019; SPNETWORK, 2016).

Concerning road networks, there is also the *OSMnx* package that lets the user download and manipulate geospatial data from OpenStreetMap. It is possible to model, project, visualize, and analyze real-world street networks and any other geospatial geometries (OSMNX, 2021).

Flashgraph is a suitable option for handling large networks when nodes and edges exceed the available main memory capacity. It is possible thanks to its intelligent I/O scheduling and its storage system, which leaves only the network vertices in main memory and the edges in SSDs (ZHENG et al., 2015).

Urban Network Analysis is another tool we can mention. It was launched by City Form Lab and is available in Rhino and ArcGIS software, allowing the analysis of spatial networks. It only encompasses a few metrics that are more specific to urban networks, such as *betweeness*, *closeness* e *straightness* (SEVTSUK; MEKONNEN, 2012).

neo4J is a graph database with native storage and processing. It comprises a graph property model, and a query language called Cypher that facilitates the understanding and use of the tool (NEO4J, 2020).

Table 3.2 - Tools for network creation and analysis.

Name	**Type**	**Language**	**Clear Spatial Support**	**Benefits/Particularities**
igraph	Library Collection	Python, C, R, Mathematica	No	Good performance and portability
networkx	Library	Python	Yes	Network construction from different data types
tidygraph	Package	R	No	Based on data frames, allows relational data manipulation, such as graph joining
stplanr	Package	R	Yes	Designed for transport planning
dodgr	Package	R	Yes	Especific for road networks
spnetwork	Package	R	Yes	Performs spatial analysis on networks
OSMnx	Package	Python	Yes	Especific for street networks based on Open Street Maps
flashgraph	Package	C++, with R bindings	No	Semi-external memory graph processing engine, optimized for a high-speed SSD array
Urban Network Analysis	Toolbox (ArcGIS and Rhino)	Python	Yes	For spatial analysis on urban street networks.
neo4J	Database	—	Yes	Graph database with native storage and processing

Table 3.2 presents a comparison of the mentioned tools. In this work, we use mainly

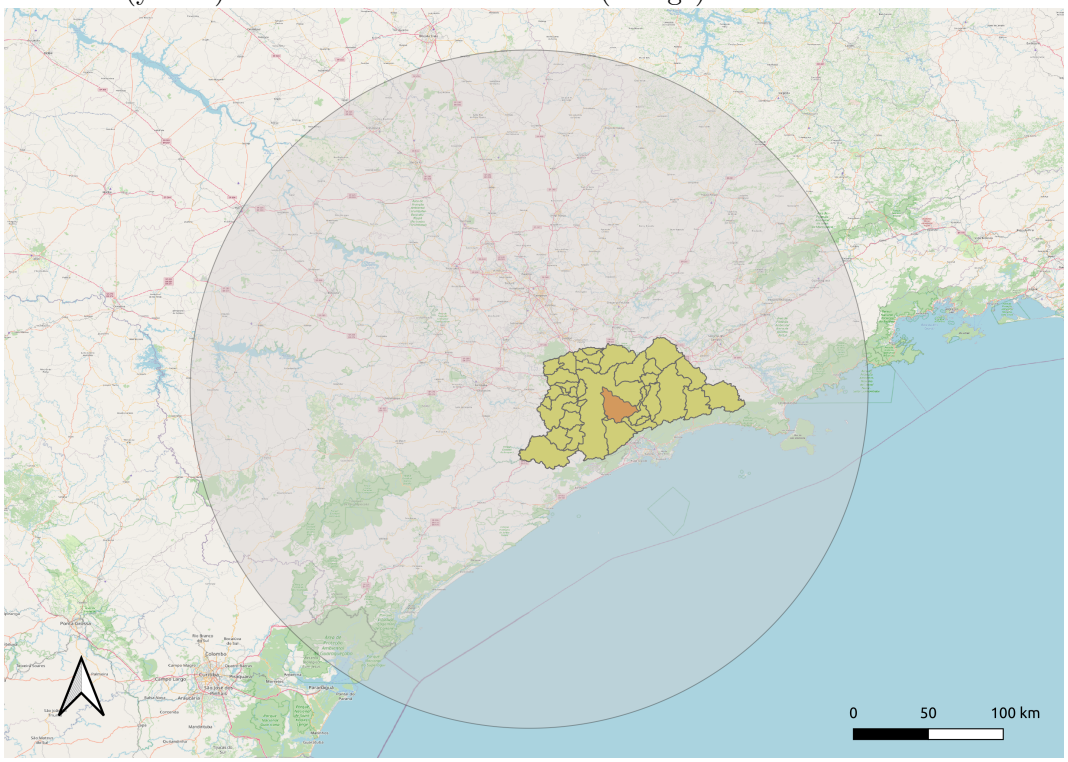
the igraph library, jointly with the Python language, due to its portability, good performance and the vast amount of functions available for analyzing networks. We also employ Networkx library in order to use some specific functions.

4 METHODOLOGY

4.1 Data

For having a high spatial and temporal resolution, weather radar data were adopted as the basis for the present research. Given the areas addressed in the case studies, which we describe in the following chapters, the radar with the most suitable coverage is situated in the municipality of São Roque (lat= $23°35'56"S$, lon= $47°5'52"W$). Its operation is in charge of the Department of Airspace Control (DECEA), and its range is 250 km in qualitative mode and 400 km in surveillance mode. Figure 4.1 presents the coverage considering the qualitative mode. The Metropolitan Area of São Paulo (MASP) is highlighted in the map to show the coverage geographically.

Figure 4.1 - Coverage of São Roque weather radar, the Metropolitan Area of São Paulo (yellow) and the Tamanduateí Basin (orange).



This weather radar performs a volumetric scan composed of azimuth scans in 15 different elevation angles, from 0.5 degrees to nearly 20 degrees. Each of these azimuth scans generates a product named PPI (Plan Position Indicator). It consists of projecting an azimuth scan onto a horizontal plane. All data are transformed from

polar coordinates into a cartesian grid. CAPPI (Constant Altitude Position Indicator) is another product derived from radar scans. It is a projection of a horizontal plane at a constant height (REDEMET, 2015). The volumetric scans provide high-resolution data in space and time domains, with approximately 1km and 10 minutes, respectively (DEPARTAMENTO DE CONTROLE DO ESPAÇO AÉREO (DECEA), 2010).

The output is a set of reflectivity values (target return echoes) in dBZ units, obtained from the reflectivity factor logarithm (Z), considering the default diameter of rain droplets. Using the Marshall-Palmer formula, reflectivity values correlate to an estimated rainfall rate (R) (MARSHALL et al., 1947). For all our case studies, we keep the values in reflectivity units (dBz) as they are available. In summary, the higher the reflectivity value, the more intense is the estimated precipitation.

4.2 (geo)graphs: geographical graphs

Spatial dependence is a fundamental physical property of many phenomena modeled through networks. It is also the case of the meteorological networks addressed in this research. They are phenomena set in the atmosphere whose geographic component is intrinsic in the formation and occurrence of events.

Santos et al. (2017) presents the concept of (geo)graphs: graphs whose nodes have a known geographic location and the edges have an intrinsic spatial dependence. Furthermore, they are objects compatible with Geographic Information Systems (GIS), a classic approach to manipulating spatial data.

This concept is one of the basis of this research for constructing precipitation networks since it includes the geographic component. Moreover, it allows data exportation in shapefile format, making it possible to conduct spatial analysis on a GIS platform.

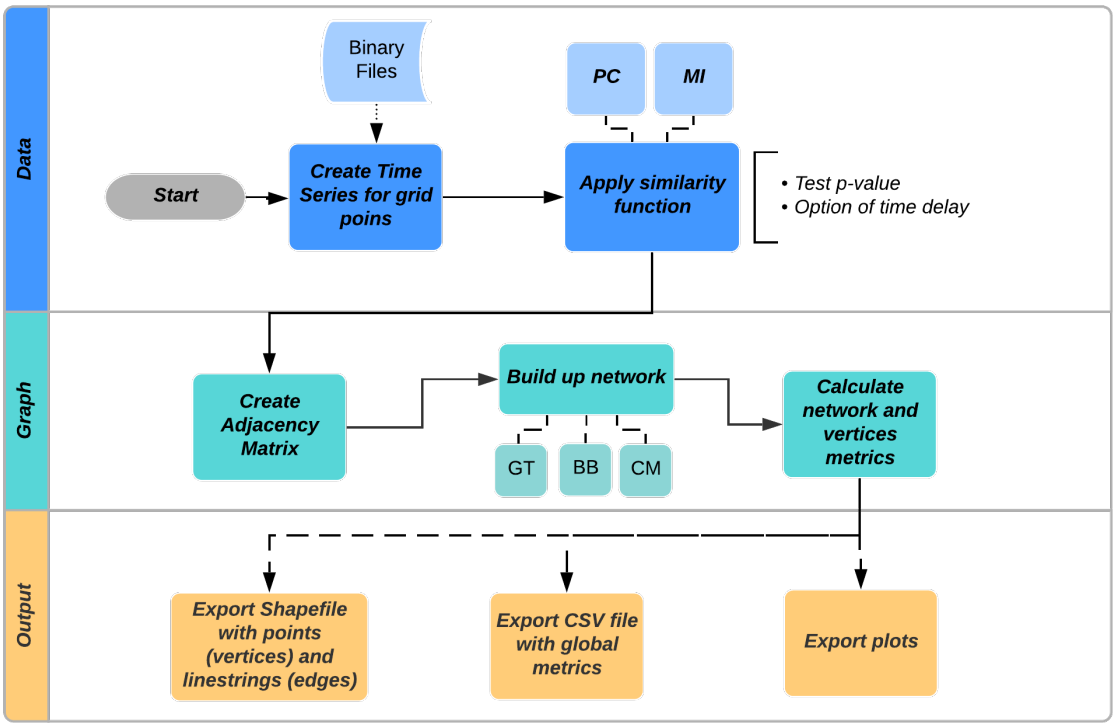
4.3 From binary data into geographical networks

We developed an application in Python, called *graph4gis (G4G)*, which reads the input data, builds up the networks, and in the end, exports the results in shapefile format. The igraph library is used for the construction of graphs and the calculation of metrics. The code was structured, at first, in 3 modules: *data, graph* and *output*. Figure 4.2 presents the flowchart of the processes involved.

The first module processes the input data by importing the radar binary files and creating a time series. Next, it calculates the similarity measure between the time

series of each pair of grid points. G4G offers two alternatives of similarity functions: Pearson Correlation (PC) and Mutual Information (MI). For Pearson Correlation, there is the option to consider a time delay when comparing the pair of time series. Only correlations with meaningful p-values are considered (significance level = 0.05).

Figure 4.2 - G4G Flowchart.



The second module deals directly with graph issues. First, it generates a weighted adjacencies matrix from the calculated similarity function. Then, it builds up the graph using one of the following criteria: Global-Threshold (GT), Backbone (BB), or Configuration Model (CM). Section 2.4 clarifies how each criterion works. As a result, each node represents a grid point, and the selected edges carry the similarity weights. In the end, this module calculates the metrics for each vertex and the global metrics of the network. Section 2.5 describes all the metrics G4G calculates.

The third module is responsible for exporting the results. It creates a shapefile with a set of points and lines, geographically representing the vertices and edges of the graph. Next, the module adds the calculated metrics for each vertex as geometry attributes in this shapefile. The application also delivers a CSV file with global network metrics. This module is furthermore responsible for all the plots that G4G

produces.

5 FIRST CASE STUDY: SPATIAL DEPENDENCE AND RELATIONS BETWEEN TOPOLOGICAL AND GEOGRAPHICAL DISTANCES

This chapter is based on the paper "Geographical complex networks applied to describe meteorological data" published in the Proceedings of the XXII Brazilian Symposium on Geoinformatics (GEOINFO). It presents our first case study, which involves analyzing the relations between topological and geographical distances and the spatial dependence inherent in the network structure. Besides, we analyze the geographical layout of the network on a watershed.

5.1 Introduction

Based on Graph Theory, the study of Complex Networks represents a relevant contribution to science as a tool to describe the structure of a wide range of complex systems in nature, such as climate events (BARABÁSI; PÓSFAI, 2016). In such a context, Complex networks have been applied to climate data analysis to identify structural patterns and teleconnections. The researchers use similarity measures such as Pearson correlation, event synchronization, or mutual information to construct the network connections. In terms of data, they are based on long time series of atmospheric variables, ranging from months to several years (TSONIS et al., 2006; BOERS et al., 2019).

Few works have been held specifically in the weather domain, dealing with short-term changes in the atmosphere and manipulating spatial and temporal high-resolution data through complex networks. One of those few examples handled precipitation data from weather radar, and they achieved significant results in community detection based on a time series of only ten days, with 1 kilometer of spatial resolution (CERON et al., 2019). The behavior of topological metrics in meteorological networks is a characteristic that remains unknown.

With this in mind, the present work aims to make some progress in the spatial analysis of metrics in meteorological networks, specifically in precipitation events.

Due to climate changes, extreme precipitation events are becoming more frequent, with several impacts on society. Finding spatial patterns of precipitation events could represent a significant advance in atmospheric science and several applications, from health geography to resilient urban mobility (SANTOS et al., 2017).

5.2 Materials and methods

5.2.1 Data

The case study presented here was held in São Paulo Metropolitan Region, specifically comprising the area of Tamanduateí basin, from January 2015. Located on the Tiete river's left margin, the Tamanduatei basin has its source in the city of Mauá. It also crosses the towns of Diadema, São Caetano do Sul, besides the eastern and central zones of São Paulo (RAMALHO, 2007).

As previously mentioned, we used weather radar time series as our base dataset due to its spatial and temporal high-resolution data. The weather radar of São Roque, described in Section 4.1, is the one that offers the best coverage.

For the present study case, we used PPI data corresponding to the first elevation level. Only reflectivity values above 20 dBz were considered as it represents an estimated rainfall rate of 1 millimeter per hour. This way, we also disregard any possible noisy data. The selected time series comprises the entire month of January of 2015 with a temporal resolution of 10 minutes, so it is composed of more than 4400 scans in time, each one of them including 783 points in space.

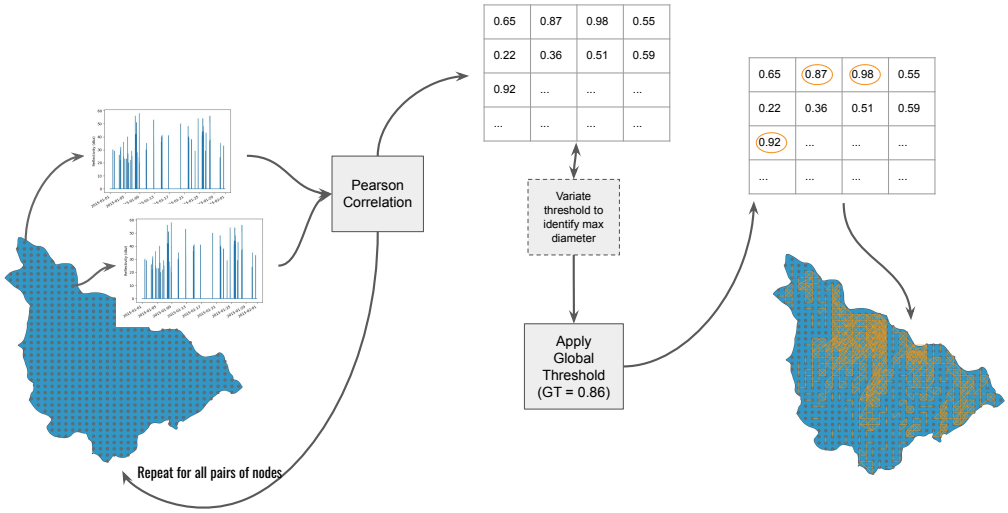
5.2.2 Network construction and analysis

Spatial embedding is a physical property inherent in many phenomena modeled through networks, including the meteorological events addressed in this research. Therefore, we use here geographical graphs, as described in Section 4.2

We developed a tool to manage the input data and construct the network considering its geographical component. We called it *Graph4GIS (G4G)*. It delivers output files with topological metrics calculated. One of these outputs is a shapefile, a file compatible with GIS platforms, allowing graph visualization in geographical space. Section 4.3 provides a complete description of *G4G* operation flow.

Pearson Correlation is the similarity measure we employ in this case study. The network is built up using the Global-Threshold (GT) criteria (details in 2.4). Figure 5.1 describes the flow employed by G4G for this case study.

Figure 5.1 - G4G flow for Case Study 1.



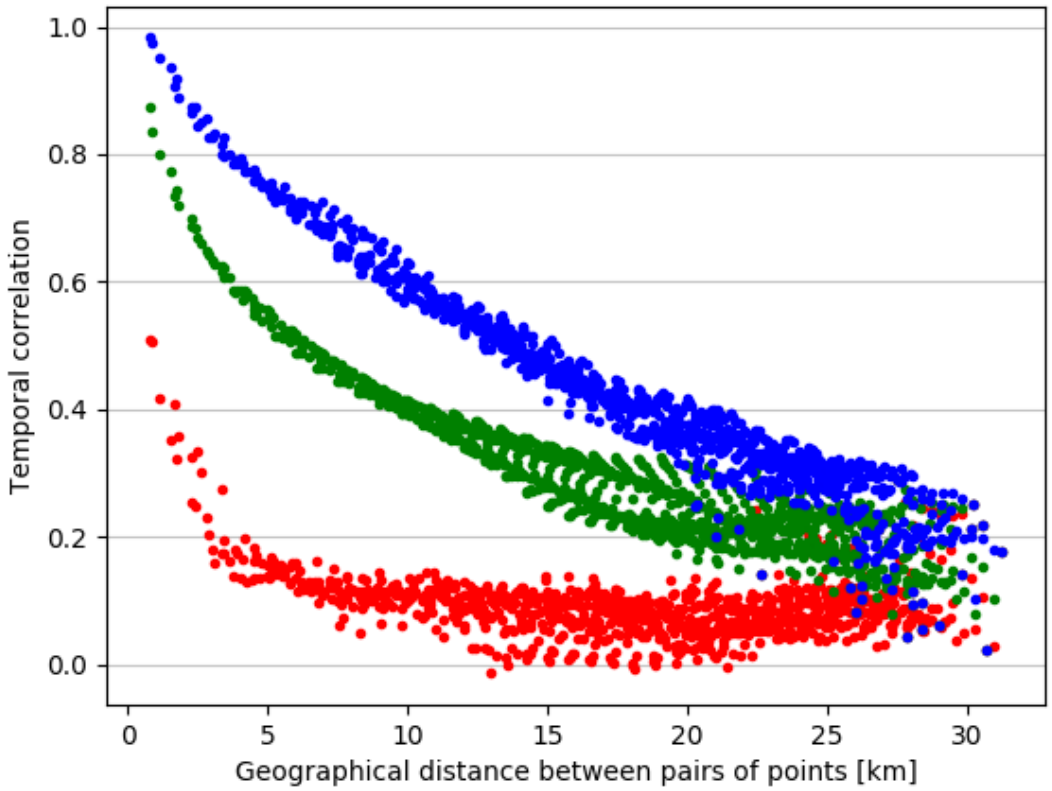
5.3 Results

Before analyzing the network built for the mentioned case study, we can observe the spatial dependence inherent in such data on Figures 5.2 and 5.3. The first one shows how the (temporal) correlation between the (time series associated with each) pairs of points is related to the geographical (euclidean) distance between them. We grouped correlation values into three categories - minimum, medium, and maximum - respectively colored in red, green, and blue. For the red group, it is possible to see a spatial dependence up to approximately 3 km. Regarding medium and maximum categories, the temporal correlation is considerably high between 1 and 10 km of distance, but we can still observe the influence of spatial dependence until 20 km.

We can also notice that the minimum correlations for the geographically nearest ten pairs of points are even higher than the maximum correlations for those more distant than 28 km. Such property is an indicator of how well-behaved the relation between temporal correlations and geographical distances is in this network structure.

The scatter plot on Figure 5.3 presents the relation between the euclidean distance and the topological distance between each pair of nodes - the network path with the shortest number of edges between those nodes. We can verify strong linearity in such relation, with a correlation coefficient (R^2) equal to 0.767 and a slope of 1.16. Such a slope value indicates that as the geographical distance increases, the impact

Figure 5.2 - Temporal Correlation versus Geographical distance between each pair of points. Correlation values are grouped into three categories - minimum, medium and maximum values - respectively coloured in red, green and blue.



is even more significant on the topological distance.

This chart also shows the longest edge in the network (2.5 km), indicated by the maximum geographical distance for the pairs of points within a topological distance of 1 edge. Therefore, there are no pairs of points directly connected at a distance greater than 2.5 km. On the other hand, there are very close nodes, geographically neighbors, but with a high topological distance, up to 12 edges.

The geographical network built up by graph4GIS is introduced in Figure 5.4. It used a threshold of 0.86, which was the critical threshold for our study case. This output allows us to visualize the structure of network connections spatially.

Figure 5.3 - Topological distance versus Geographical (euclidean) distance.

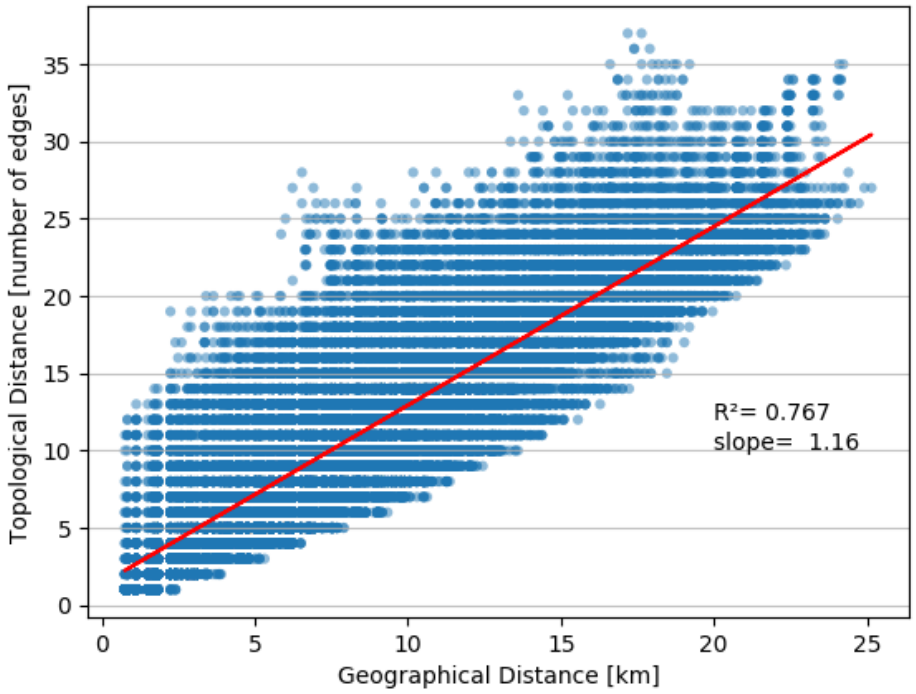
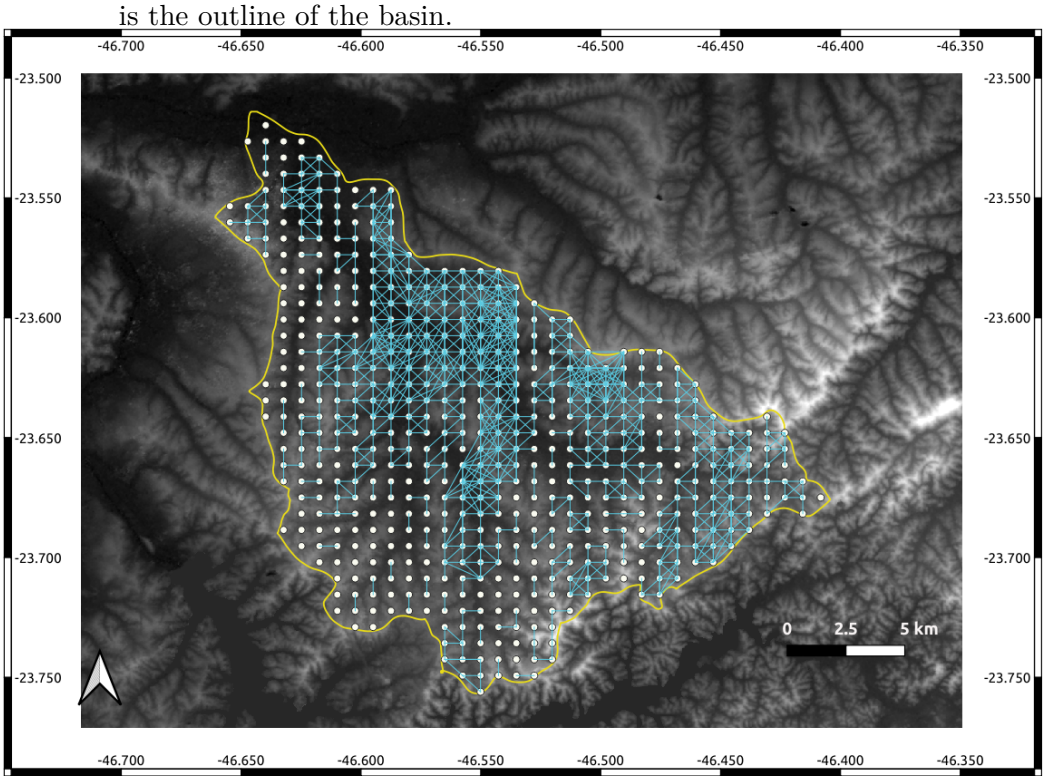


Figure 5.4 - Geographical network for Tamanduateí Basin. The white points represent the nodes, the blue segments are the edges of the network, and the yellow border is the outline of the basin.



5.4 Final considerations

This work applied Complex Networks in the study of meteorological networks, aiming to explore topological metrics' behavior in such a context. This paper introduced some spatial analysis of the system's topological structure based on precipitation time series.

As a result, we could identify the spatial dependence of temporal correlations, such as the linearity in the relation between the topological and geographical distances between different pairs of points in a hydrological basin. We also verified some peculiarities in the network, such as the maximum geographical length of an edge (2.5 km) and a high maximum topological distance between neighboring nodes (11 edges on the shortest path between nodes close 1km to each other).

In future works, we would like to analyze datasets for specific meteorological processes to identify spacial and topological signatures. Besides, we intend to approach larger study areas, including the entire São Paulo Metropolitan Region, and other graph measures, such as degree, clustering coefficient, betweenness, and diameter.

6 SECOND CASE STUDY: BUILDING UP DIFFERENT NETWORKS IN A WATERSHED

We based this chapter on the paper "Global-threshold and backbone high-resolution weather radar networks are significantly complementary in a watershed" submitted to *Computers & Geosciences*. It presents the second case study, which compares different similarity measures and criteria for building up networks. We analyze the structural patterns of the various generated networks.

6.1 Introduction

Complex networks have been widely applied in the study of several complex systems in nature and society (BARABÁSI; PÓSFAI, 2016). In climate, they are used as an alternative tool for investigating climate dynamics (FERREIRA et al., 2021). Based on long-term events, such studies analyze atmospheric datasets in long time series that range from months to several years (TSONIS et al., 2006; BOERS et al., 2014).

Boers et al. (2019) revealed a global pattern of extreme-rainfall teleconnections using an event synchronization method. They employed a satellite-derived rainfall dataset in a daily temporal resolution for almost 20 years. Tsonis et al. (2006) identified super-nodes associated with teleconnection patterns based on reanalysis data. Such a dataset comprehended over fifty years in a temporal resolution of one month. Their findings suggest that the organization of teleconnections is related to the stability of the climate system.

Differently, the weather, which deals with short-term changes in the atmosphere, has been little exploited in Network Science. Ceron et al. (2019) published one of the few studies within this context, using spatial and temporal high-resolution data from weather radar to detect community structures. Jorge et al. (2020) also worked with weather radar networks, analyzing the relation between topological and geographical distances.

In this work, we build (geo)graphs (geographical networks (SANTOS et al., 2017)) based on weather radar data, creating connections between points on the geographical space in a watershed. We use different similarity measures: Pearson Correlation (PC) coefficient and the Mutual Information (MI) index (KRASKOV et al., 2004). The former captures linear correspondences between the series, while the latter also recognizes nonlinear relations. Besides, we follow two criteria: a global threshold to connect similar time series and a local threshold criterion to extract the network

backbone. Interestingly, both criteria generate significantly complementary network structures. We compare them, taking into account the geographical context. Our findings also show a statistically significant linear relationship between topological and geographical distances.

6.2 Material and methods

6.2.1 Weather radar dataset and the Tamanduateí basin

Our study area focuses on the Tamanduateí river's basin in the Metropolitan Area of São Paulo, Brazil. Situated on the Tiete river's left margin, the Tamanduateí basin has its source in Mauá. It also crosses the cities of Diadema, São Caetano do Sul, and the eastern and central zones of São Paulo (RAMALHO, 2007). It is one of the basins with the highest number of extreme rainfall events in the city of São Paulo (COELHO, 2016).

We employ the Shuttle Radar Topography Mission (SRTM) Digital Elevation Model data to define the watershed contour. A *watershed* is a land area that drains rainfalls and streams to a common outlet such as a bay mouth or any point along a stream channel. By employing a digital elevation model, it is possible to identify the drainage system based on the land's topography.

We analyze the precipitation series from January 2015, with data from a weather radar located in the city of São Roque, described in Section 4.1, which is 60 km distant from our study area.

We use only the first azimuth scan (PPI) and values as they are available in dBZ units. Such data are arranged in a cartesian grid after being converted from polar coordinates. From this initial grid, we select only the points strictly inside the limits of the Tamanduateí basin and values that correspond to a precipitation rate above 1 millimeter per hour.

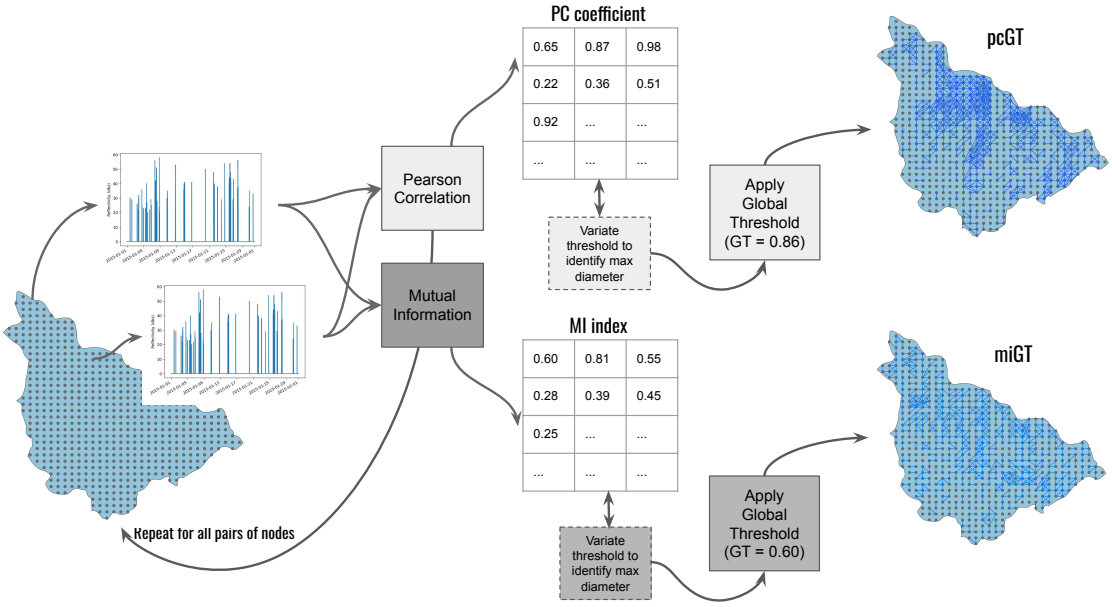
6.2.2 The construction and analysis of the geographical networks

The nodes of our graph are the grid points of the weather radar scan, which are inside the Tamanduateí basin area. We preserve its spatial location as an attribute. As a result, if we plot it over a map, we have 587 nodes, one by square kilometer, inside the basin boundaries. The edges are based on similarities between the corresponding time series of each pair of vertices. We use and compare two similarity functions in this work: the Pearson Correlation (PC) coefficient and the Mutual Information

(MI). Both are described in Section 2.3.

Once all edges have associated similarity measures, one must define which must remain in the network. We use and compare two criteria for that purpose: a global threshold (GT) to connect similar time series and a local strategy to extract the network backbone (BB). In the first one, the edges remain for those pairs whose similarity exceeds a global threshold, defined as the point of maximum diameter of our network (SANTOS et al., 2019). Distinctly, the BB criterion consists of using each node's weight fluctuations to select the edges to be preserved. Both GT and BB criteria are explained with more details in Section 2.4.

Figure 6.1 - G4G flow for Case Study 2 — pcGT and miGT networks.

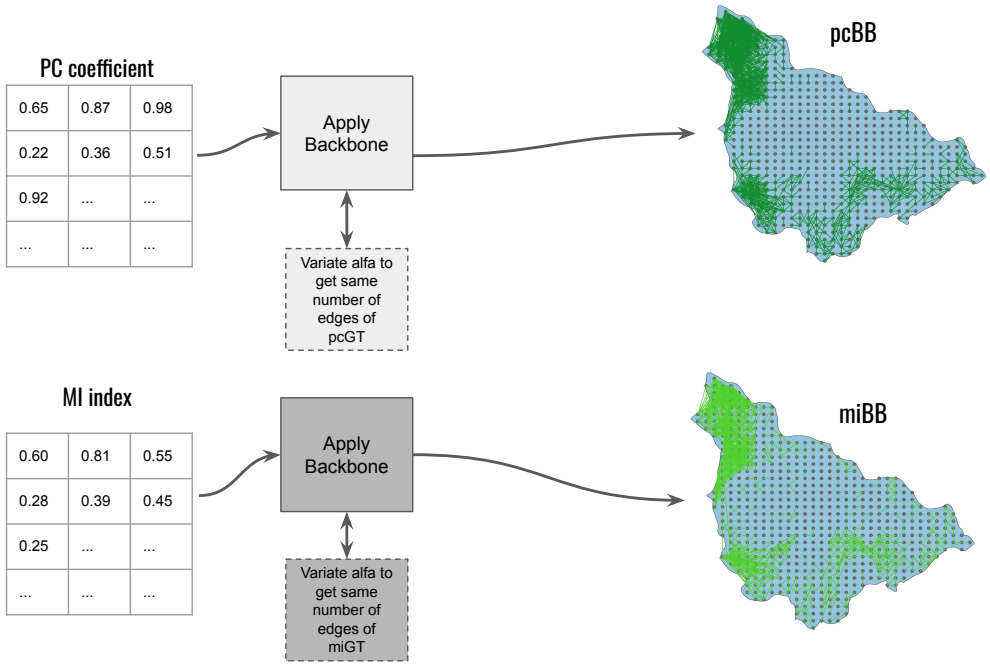


Combining the two similarity metrics, PC and MI, with the two network-building criteria, GT and BB, we end up with four network structures: pcGT, pcBB, miGT, and miBB. We build the pcGT with a global threshold of 0.86, which results in a graph with 587 vertices and 1270 edges. Figure 6.1 describes the flow to build up pcGT and miGT. Then, for the pcBB, we adjust the value of α so that the resulting network has approximately the same number of edges as the pcGT.

When applying the Mutual Information in our dataset, the range of weights is lower than the obtained with the Pearson Correlation coefficient. The threshold of the

maximum diameter is 0.6, and the miGT has fewer edges (964) than the pcGT. To build the miBB, we adopt the same idea as before, adjusting the value of α in order to achieve a resulting network with the same number of edges as the miGT. The pcBB and the miBB are built up according to the flow described in Figure 6.2.

Figure 6.2 - G4G flow for Case Study 2 — pcBB and miBB networks.

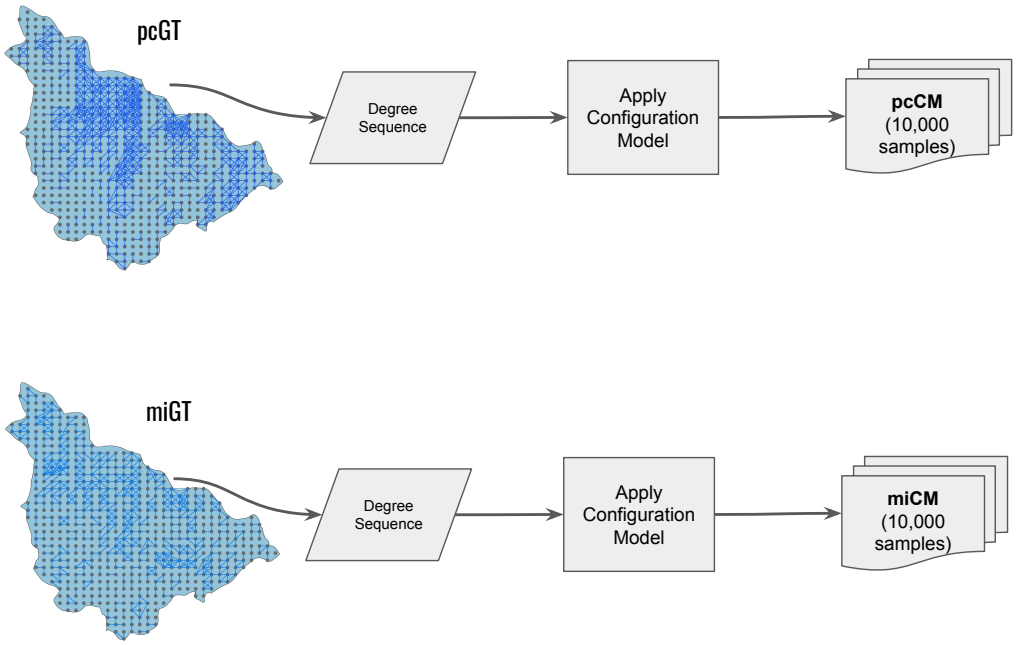


We compare the mentioned networks with the Configuration model (CM) (NEWMAN, 2003), a null model based on random connections. The Networkx library provides an implementation that generates pseudographs by randomly assigning edges to match a given degree sequence. We generate a set of ten thousand samples of pseudographs based on pcGT's degree sequence and another ten thousand samples based on miGT's. We refer to these sets as pcCM and miCM, respectively. Figure 6.3 presents G4G flow concerning the networks pcCM and miCM.

For all the constructed network models, we measure and analyze the network metrics: average shortest path ($\langle l \rangle$), average of the local clustering coefficients ($\langle c \rangle$), diameter (D), heterogeneity parameter (κ), number of components (NC), the size of the giant component (GC) and the number of singletons (ST). The first one refers

to the average of the shortest paths that link every pair of vertices. The local clustering coefficient indicates how connected a node's neighbors are to each other. The diameter is defined as the longest shortest path of a network. The κ is the ratio between the squared degree and the square of the average degree ($\kappa = \langle k^2 \rangle / \langle k \rangle^2$). Large κ means heavy-tailed degree distributions, while $\kappa \approx 1$ approximates to random networks (Poisson distribution). The NC is the number of isolated groups of nodes (components), GC is the size of the largest component, and ST is the number of components with a single node (BARABÁSI; PÓSFAI, 2016).

Figure 6.3 - G4G flow for Case Study 2 — pcCM and miCM networks.



Furthermore, we analyze if each network structure could be statistically classified as a small-world network by comparing the $\langle c \rangle$ and $\langle l \rangle$ between the original network and an equivalent random one with the Erdős and Rényi model (BARABÁSI; PÓSFAI, 2016) — the same number of nodes and edges, but not necessarily preserving the degree distribution. The small-world phenomenon implies a short distance between two randomly chosen nodes in a network. Section 2.6 describes how this comparison can be done.

6.3 Results and discussion

Table 6.1 presents the weight (PC coefficient or MI index) range and topological metrics for each network model. Concerning the CM models, the values are the average numbers since they correspond to several realizations. For the same reason, we could not measure the values of NC, GC, and ST for them. The last two columns contain the $\langle l \rangle$ and $\langle c \rangle$ metrics for the equivalent random network (Erdős and Rényi model) in each case (same number of vertices and edges, without preserving the same degree distribution).

Table 6.1 - Topological properties of each network.

Network	L	Weight	$\langle l \rangle$	$\langle c \rangle$	D	κ	NC	GC	ST	$\langle l_{\text{rand}} \rangle$	$\langle c_{\text{rand}} \rangle$
pcGT	1270	0.86-0.98	8.93	0.536	37	1.775	125	349	85	4.35	0.007
pcBB	1269	0.27-0.95	4.42	0.225	19	3.263	237	161	218	4.34	0.007
pcCM	1270	0.13-0.96	3.89	0.017	8.66	1.775	–	–	–	–	–
miGT	964	0.6-0.85	10.38	0.474	49	1.422	82	305	33	5.36	0.005
miBB	964	0.18-0.69	3.88	0.159	19	2.962	202	152	152	5.36	0.005
miCM	964	0.04-0.74	5.06	0.007	11.5	1.422	–	–	–	–	–

As Table 6.1 shows, the average shortest path ($\langle l \rangle$) and the diameter (D) are higher when using a global threshold. The same occurs with the clustering coefficient ($\langle c \rangle$), which increases as the network preserves similar connections. Differently, the heterogeneity parameter (κ) is higher when using the BB criterion, showing that the degree distribution tends to be less homogeneous when using a backbone strategy. It is also possible to notice that we generate less fragmented networks using the GT approach. The giant components (GC) of pcGT and miGT are over twice the size of pcBB's and miBB's, and the number of connected components (NC) and singletons (ST) are considerably smaller in both GT networks.

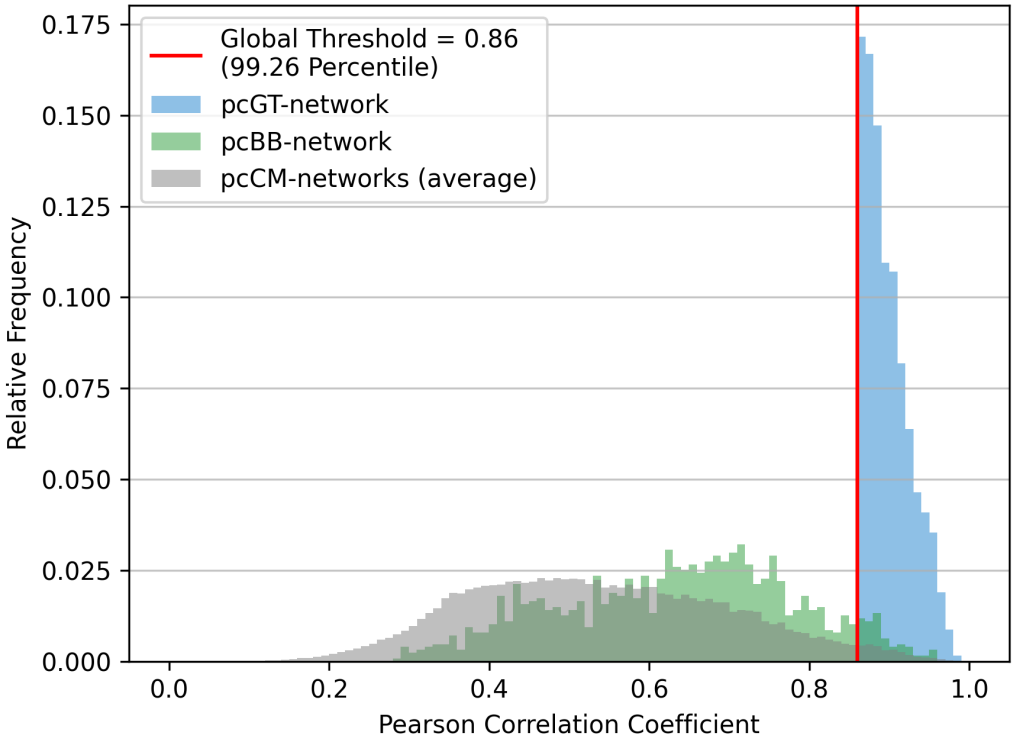
Regarding the CM networks, pcCM and miCM, the values of $\langle c \rangle$ and D are the lowest among all the network models, which is typical behavior of random networks. Their κ are precisely the same as its correspondent GT network (pcCM=pcGT, miCM=miGT) since the Configuration Model preserves the degree distribution.

When it comes to analyzing the small-world property, only the miBB fulfills the requirements. It has a lower $\langle l \rangle$ (3.88) and a higher $\langle c \rangle$ (0.159) when compared with its equivalent random network ($\langle l_{\text{rand}} \rangle = 5.36$, $\langle c_{\text{rand}} \rangle = 0.005$). The miCM could be another example of an equivalent random network, but in this case, considering the average of ten thousand samples and keeping the degree distribution of the original

network. In this scenario, using miCM as a comparison base, we also have $\langle l_{\text{rand}} \rangle$ and $\langle c_{\text{rand}} \rangle$ that satisfies the conditions to classify the miBB as being statistically a small-world network.

Figures 6.4 and 6.5 contain the weight histograms of the investigated networks. The first figure contains those using Pearson Correlation, and the second one refers to the networks with Mutual Information. By definition, both GT networks (pcGT and miGT) present weights that are higher than or equal to the global threshold. Contrastingly, the pcBB and the miBB networks generate flatter distributions, as the backbone method also considers the nodes' degree and strength besides the edge's weight. The number of shared edges between pcGT and pcBB is 78 (6.14% approximately), and between miGT and miBB is 71 (7.36%).

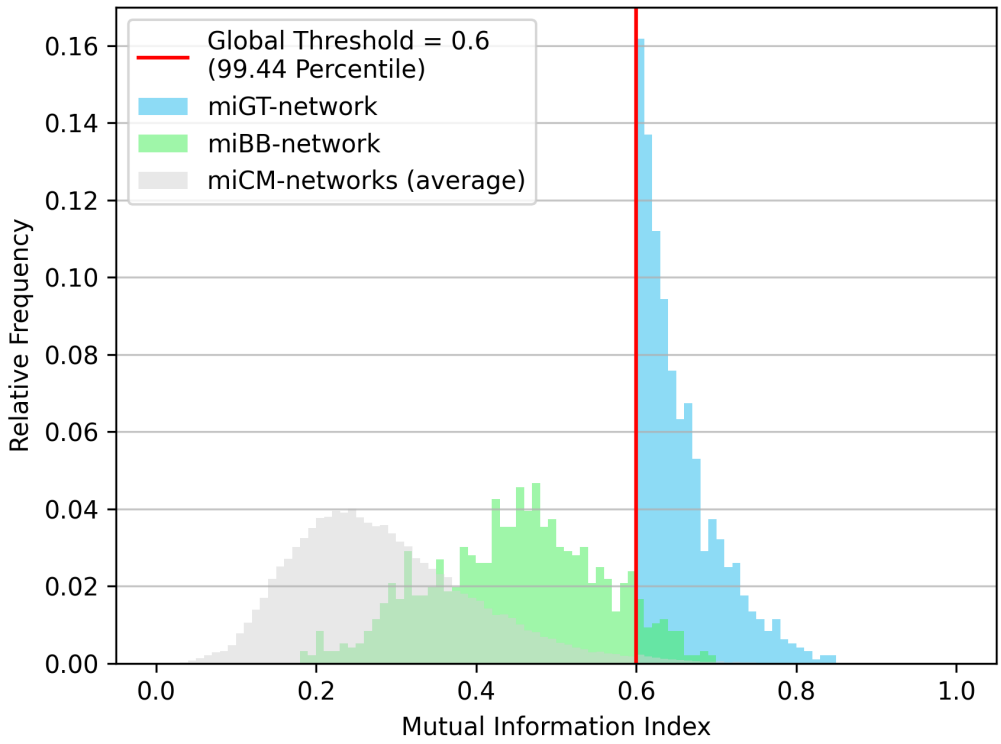
Figure 6.4 - Weight distribution for PC networks.



The CM networks present distribution with a higher frequency on lower weights when compared to their corresponding BB or GT. When comparing the set of PC networks with the set of MI, one notices that the shapes of the distributions are

similar, with a slight difference in the relative frequencies. MIs have more edges with lower weights, naturally expected since the MI index range is smaller than the PC coefficient range.

Figure 6.5 - Weight distribution for MI networks.

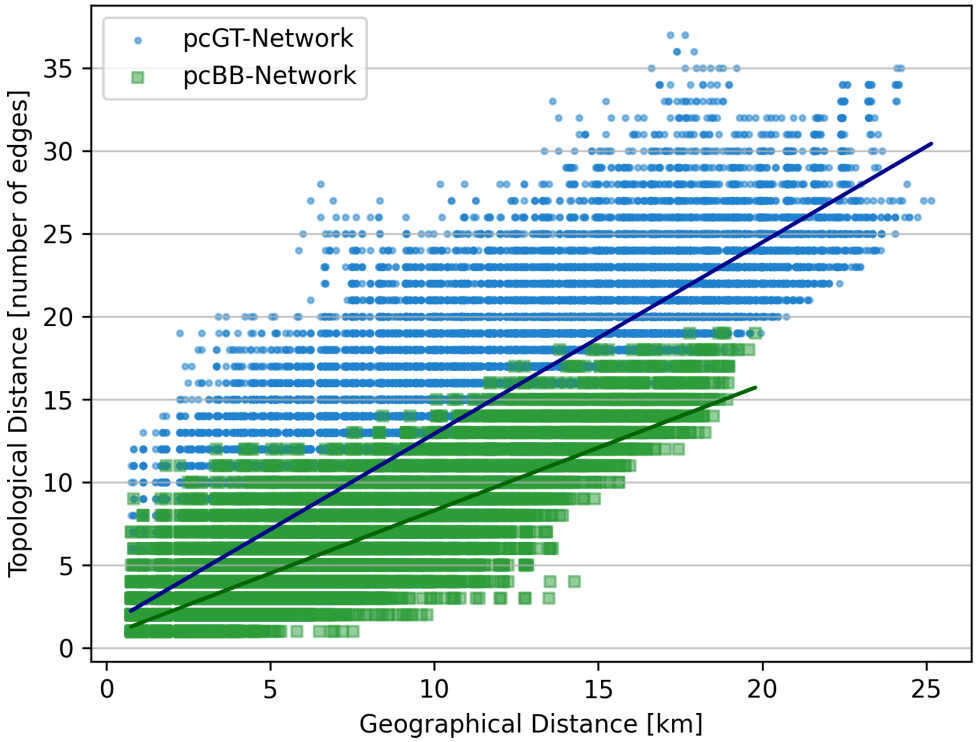


Figures 6.6 and 6.7 introduce the scatter plots of the topological distance *versus* the geographical (euclidean) distance - each point represents a pair of nodes. It compares the spatial dependence between GT and BB networks, both based on PC and MI. In case of both the PC networks, there is a significant linear relationship: $R2 = 0.77$ with a slope of 1.15 (p-value $< E-7$) for the pcGT, and $R2 = 0.68$ with a slope of 0.76 (p-value $< E-7$) for the pcBB. Concerning the MI ones, only miGT presents a high $R2$ (0.79, with slope = 1.95), whereas miBB has an $R2$ of only 0.20 (slope = 0.24).

The pcBB shows longer edges (topological distance = 1) in geographical space when compared to the pcGT. The pcBB presents an average geographical length of 2.14, and a maximum geographical length of 7.52, while pcGT has an average geographical

length of 1.06, and a maximum geographical length equals 2.40. The pcBB-network contains nodes whose time series are highly correlated to only a few others. However, some highly connected nodes (hubs) are also present due to their combination of high degree and links with lower weights, which allows long-range connections as well (geographical length > 5 km).

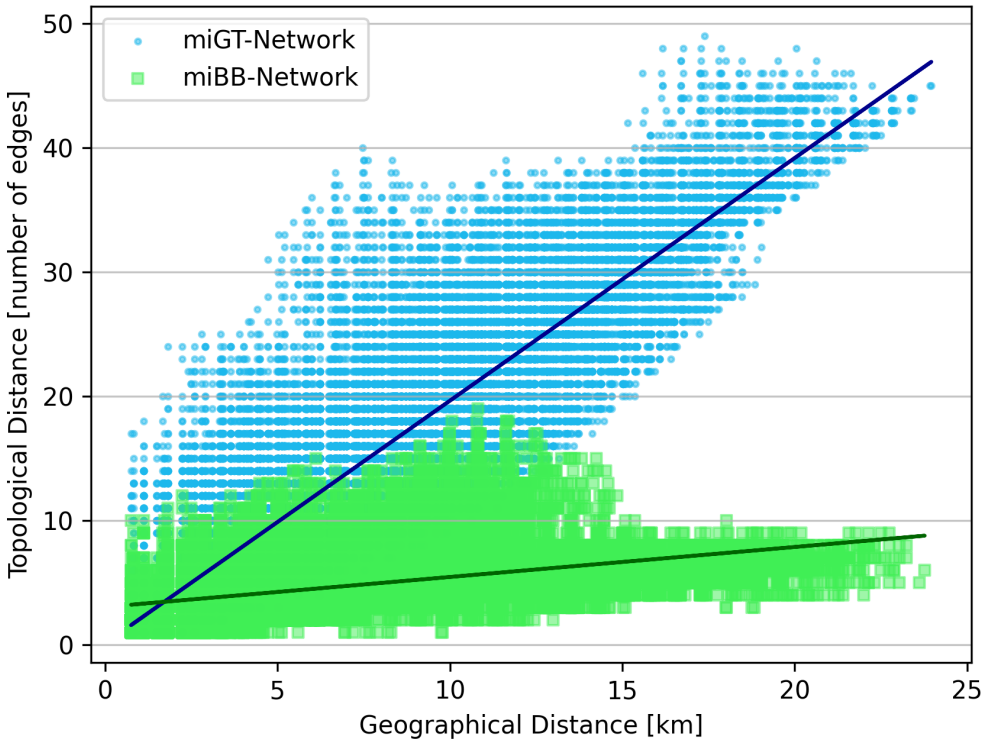
Figure 6.6 - Topological *versus* Geographical distance for each pair of nodes comparing pcGT *versus* pcBB — auxiliar lines representing linear regression, for pcGT (blue) and pcBB (green).



The behavior is similar when comparing the topological and geographical distances of the MI networks. The BB criterion also presents longer edges (topological distance $= 1$), geographically speaking, than the GT criterion. The miBB shows an average geographical length of 2.07 and a maximum geographical length of 9.78, whereas the miGT has an average geographical length of 0.94 and a maximum geographical length of 2.60. One can notice that the miBB-network has long-distance edges (almost 10 km) that meaningfully reduce the topological distances inside the network and contribute to its small-world phenomena. Distances of almost 25 km are reached

with less than 10 edges.

Figure 6.7 - Topological *versus* Geographical distance for each pair of nodes comparing miGT *versus* miBB — auxiliar lines representing linear regression, for miGT (blue) and miBB (green).



On the other hand, there are nodes in the pcBB and the miBB so close in geographical space ($\approx$ 1km) but relatively far in topological space ($>$ 5 edges). This situation is even more apparent for the pcGT and the miGT networks. As BB-networks' connected components are smaller than those from the GT-networks', topological distances are smaller for the BB-networks than for the GT-networks. In terms of geographical space, the miBB presents a greater reach than the pcBB because of its long edges.

One can visually verify the spatial structure of the PC and the MI networks in Figures 6.8 and 6.9. The intersection between GT and BB structures, for both similarity metrics, is represented by the edges in red. The GT and the BB networks are significantly complementary in the studied watershed. A watershed represents

the set of points on the space with a standard outlet for surface runoff. In the background, SRTM altimetric data is employed and the lower a cell, the darker it is. The structure of PC and MI networks are very similar. The BB networks surround the watershed, mainly in the higher part of it, southwest, and around the outlet. Oppositely, the GT networks are mainly on the central watershed area, connecting cells in a region with no high difference of altimetry and high correlation in rainfall time series. When comparing the PC with the MI networks, there are slight observable differences in the map. The miGT is visually less clustered than the pcGT, confirming the average clustering coefficient values in Table 6.1 (pcGT's $\langle c \rangle = 0.536$, miGT's $\langle c \rangle = 0.474$). The miBB has long edges on the west boundary of the watershed that probably contributes to a lower average shortest path, and consequently, to the small-world effect.

Figure 6.8 - Networks on the watershed (contour in yellow): pcGT (blue), pcBB (green), and shared edges (red). In the background, SRTM altimetric data (the lower a cell, the darker it is).

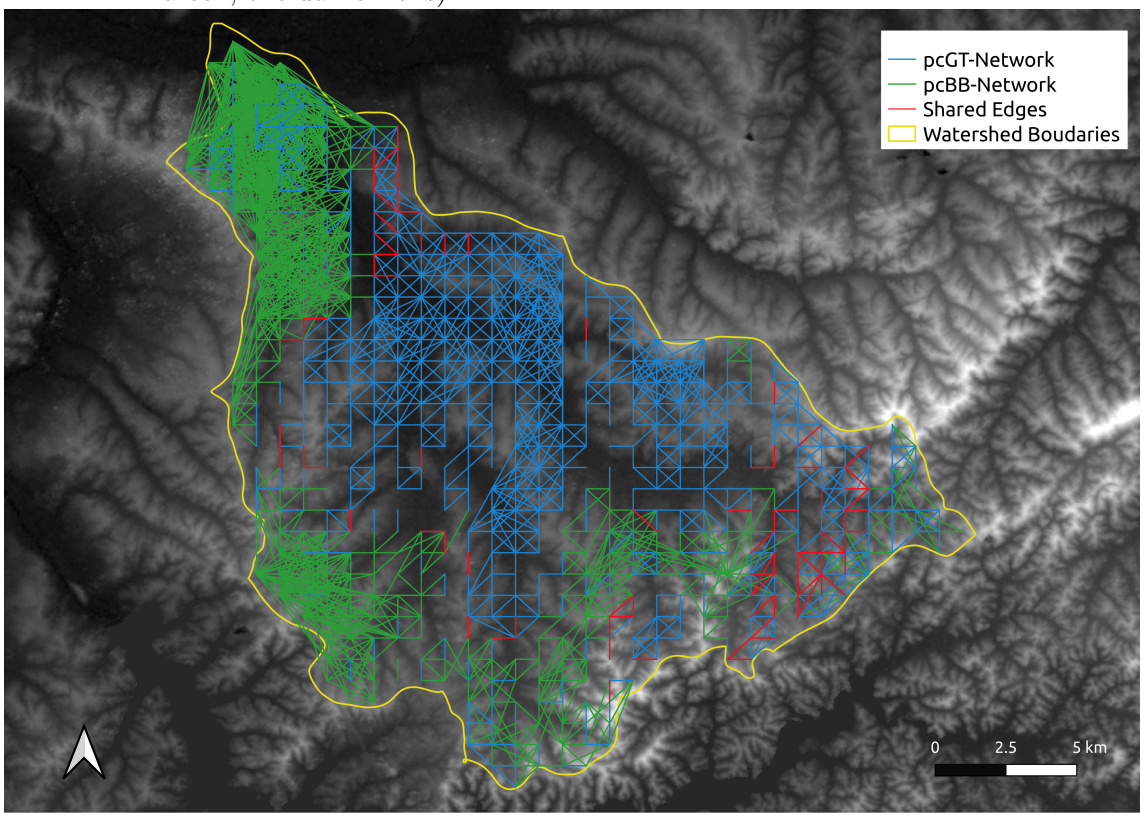
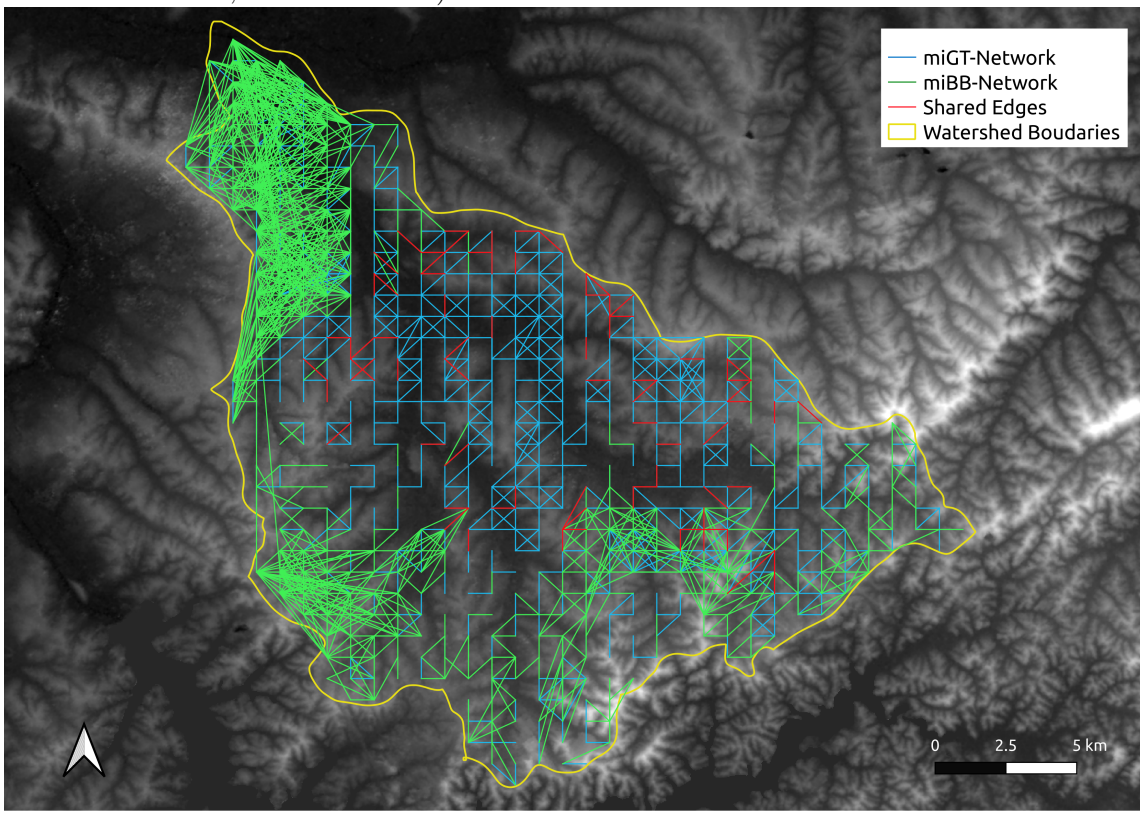


Figure 6.9 - Networks on the watershed (contour in yellow): miGT (blue), miBB (green), and shared edges (red). In the background, SRTM altimetric data (the lower a cell, the darker it is).



6.4 Final considerations

This work presented different structures of geographical networks based on weather radar data, using two similarity measures: the Pearson Correlation (PC) coefficient and the Mutual Information (MI) index. Furthermore, we employed distinct criteria to create its connections: a Global-Threshold (GT), a Backbone (BB), and a Configuration Model (CM).

The scatter plot of the topological distance *versus* the geographical (euclidean) distance revealed a statistically significant linear relationship for the pcGT and miGT networks and even for the pcBB one. The miBB was an exception presenting a low average shortest path for distant geographical points. Furthermore, our analysis showed that it could be considered a small-world network from a statistical point of view.

It is well known that the GT criterion returns a network linking nodes with very

similar behaviors (regarding its time series). On the other hand, the BB criterion provides a network linking nodes with a higher level of heterogeneity depending on the relation between the link's weight and node's strength (MENCZER et al., 2020). This work presented a geographical analysis for the different network structures in the context of a watershed: the GT networks are in the central area, close to the main rivers, while the BB networks surround the watershed and dominate cells close to the outlet. Using both PC and MI, the number of shared edges between GT and BB is only around 7% of the total number of edges, showing a significant complementarity.

As future work, we intend to reproduce the analysis in several other watersheds, from mountainous regions to floodplains, looking for spatial signatures for the different networks in the different landscapes.

7 THIRD CASE STUDY: ANALYSIS OF PRECIPITATION EVENTS AND RELATIONS BETWEEN NETWORK METRICS AND METEOROLOGICAL PROPERTIES

This chapter presents the third case study, which interprets complex network metrics in the weather context. We analyze the relations between meteorological properties and network metrics based on a set of precipitation events. The content of this chapter will be submitted to a journal to be defined.

7.1 Introduction

Recently, several works have used complex networks to support analyzing complex systems, such as the climate. By using networks, the researchers could identify teleconnection patterns and analyze the structure of climatic events. By dealing with climate, they have employed long time series of atmospheric datasets (TSONIS et al., 2006; BOERS et al., 2014).

The weather, otherwise, is related to short-term changes in the atmosphere, dealing with variables in high resolution both spatially and temporally. Very few works have explored meteorological events in network science. Ceron et al. (2019) is one of the few studies within this context handling high-resolution precipitation data from weather radar. With a dataset of only ten days, they could find community structures compatible with the land use/cover. Jorge et al. (2020) also worked with precipitation networks concerning the weather scale. They analyzed the relation between topological and geographical distances. However, none of these works have handled networks related to precipitation events.

In the present work, we build geographical networks based on precipitation events using weather radar data. These events were tracked by *Tathu* (Tracking and Analysis of Thunderstorms) from January to March 2019, focusing on the Metropolitan Area of São Paulo (MASP). In the context of this set of events, we analyze the relations between the meteorological properties and the topological metrics of the correspondent networks. Our findings show significant correlations and some particularities when analyzing specific events groups.

7.2 Material and methods

7.2.1 Data

For our study case, we used data from a weather radar situated in the city of São Roque, whose coverage includes the entire MASP. More details about this weather radar are in Section 4.1. For this work, we use the CAPPI product at the height of 3 km, which is the most used to identify meteorological systems, avoiding altitude changing and ground echoes problems. The values are used in reflectivity units (dBZ), as they are supplied by the product.

Using such data, we analyze precipitation events that occurred from January to March of 2019. To identify the events, both spatially and temporally, we used *Tathu* software, a computational tool for automatic tracking and forecasting the life cycle of weather systems. It uses techniques of image processing, geoprocessing, and spatial database. Developed as a Python package, the software architecture allows an efficient extension of functionalities and the use of different types of environmental data (UBA; GALANTE, 2021). It can identify and track events using input from satellite or weather radar data. Based on Tathu's results, we could extract the following meteorological properties related to each identified event:

- **id**: identification key that Tathu generates for each event;
- **start**: day/time when the event started;
- **end**: day/time when the event finished;
- **duration**: event duration in hours and minutes;
- **peak-lat**: latitude of the event centroid at the peak time;
- **peak-lon**: longitude of the event centroid at the peak time;
- **peak-time**: time at which the event reaches its peak with the highest reflectivity values;
- **peak-reflect_max**: value of the point of maximum reflectivity at the peak time;
- **peak-reflect_avg**: spatial average of reflectivity values at the peak time;
- **peak-area**: area of the event at the peak time (in km^2);

- **peak-area_px**: area of the event at the peak time (in number of pixels);
- **avg-speed**: temporal average of the event speed;
- **max-speed**: maximum speed identified during the event lifetime;
- **avg-area**: temporal average of the event area;
- **max-area**: maximum area identified during the event lifetime;
- **avg-reflect_avg**: the temporal average calculated over the spatial average of reflectivity values at each time step;
- **avg-reflect_max**: the temporal average calculated over the maximum reflectivity value identified at each time step;
- **delta-reflect**: the difference between *avg-reflect_max* and *avg-reflect_avg*;

7.2.2 Study area

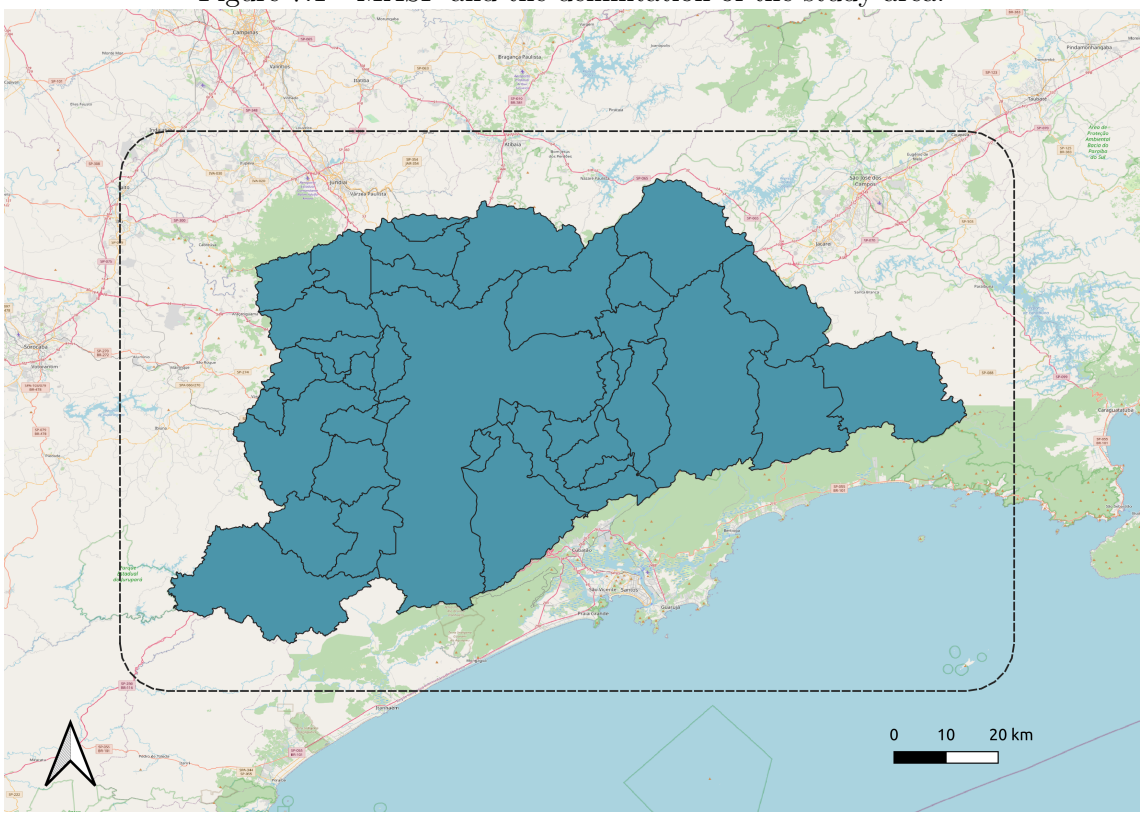
Our study area in this work is the metropolitan area of São Paulo (MASP). With 39 municipalities and more than 19 million inhabitants, it is the most significant metropolitan region of Brazil (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, 2011). Due to fast urban growth, MASP has undergone climatic changes over the past decades. As a result of these changes, temperature and precipitation show a tendency to increase, mainly since the mid's 70s (LIMA; RUEDA, 2018).

From the events database we produced using Tathu, we select those which occur inside the MASP bounding box, adding a buffer of 10 km. Figure 7.1 illustrates the delimitation of our study area.

7.2.3 Precipitation event networks

Among the events tracked by Tathu inside our study area, we filter those whose duration is at least 1 hour and 40 minutes and at most 20 hours. The lower threshold guarantees at least 10 time steps to calculate correlations later. The upper threshold is to avoid huge events with a high computational cost. After applying these filters, we end up with a sample of 383 events. For each one of these events, we build a correspondent geographical network. To do that, for each event, we select the weather radar dataset accordingly to its duration, using its start and end time as delimiters.

Figure 7.1 - MASP and the delimitation of the study area.



In the next step, such dataset is used as input to G4G (described in Section 4.3). As previously mentioned, G4G converts each grid point of the dataset, inside the selected study area, into a network node carrying an attribute of geographical coordinate. Then, the software selects the time series associated with every grid point and binds it to the correspondent node. Nodes with a completely zeroed time series are discarded from the network. Figure 7.2 presents G4G flow to build the geographical networks from this case study.

We adopt Pearson Correlation as the similarity function for the present case study, adding an option of time delay in the calculations. This delay ranges from 0 to 30 minutes, and we keep the delay that maximizes the correlation for each pair of nodes. In the end, we have two filled matrices: one with the delays and the other one with the respective correlations (weight matrix).

The Global Threshold (GT) criterion is applied to the weight matrix to select the most relevant weights. Then, the software builds an edge for the correspondent pairs of nodes. At the end of G4G processing, we have a geographical network for

the precipitation event with a group of network metrics associated (described in Section 2.5).

Figure 7.2 - G4G flow for Case Study 3.

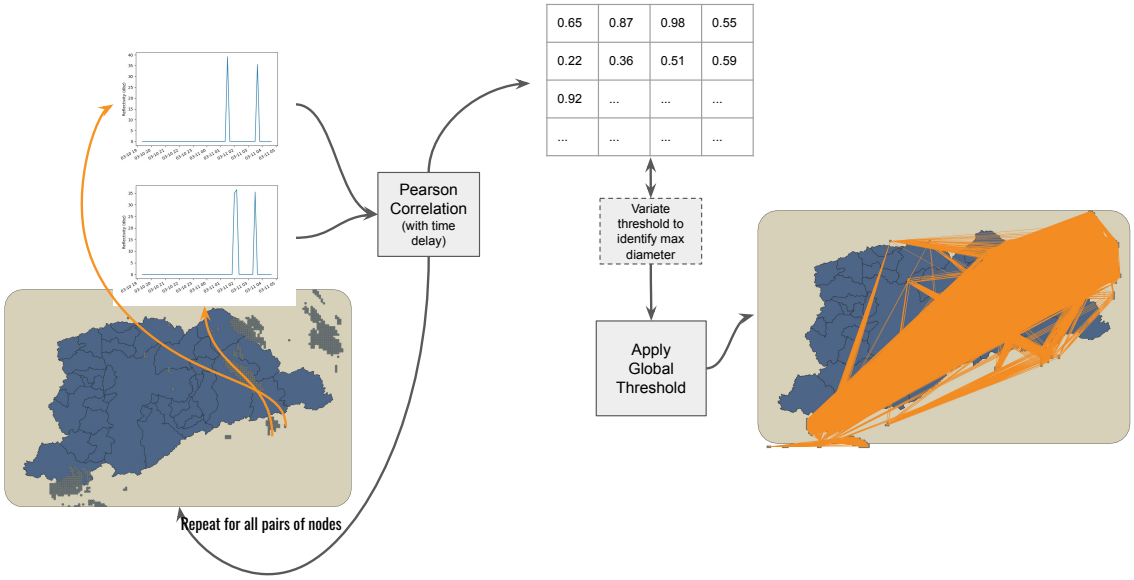
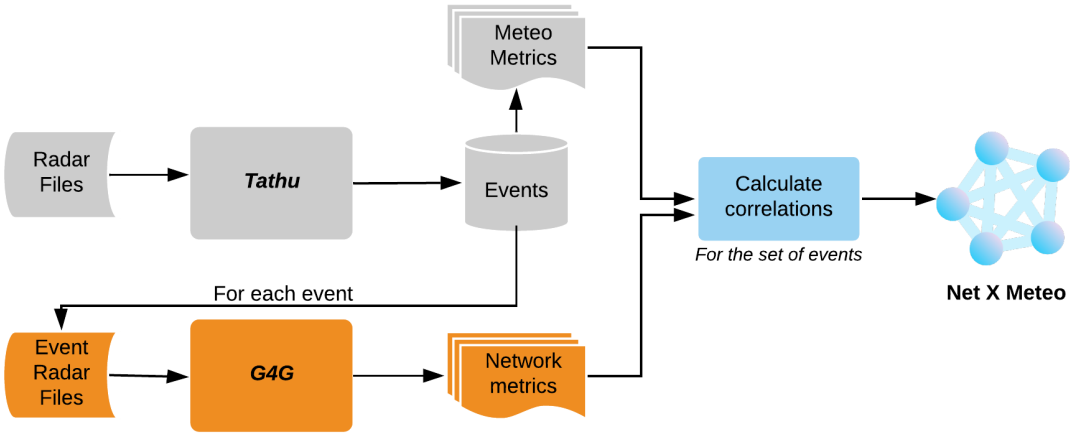


Figure 7.3 - Flow to identify correlations between network and meteorological metrics.



Once we have the network metrics for the whole set of selected events, we can try to

identify relations between these metrics and the meteorological properties. Figure 7.3 presents the flow we implemented for that purpose. After Tathu processing, it produces a database of events. Based on that, we extract a set of physical properties that characterizes the events. We call these properties meteorological ("meteo") metrics. For each event tracked by Tathu, G4G builds the correspondent network and calculates its metrics. A Person Correlation is performed for every pair meteo-network metric, discarding those correlations with a p-value not statistically significant. The result is a graph representing these relations between network and meteorological metrics.

7.3 Results and discussion

Figure 7.4 shows the resulting graph with the meteo-network correlations for the entire set of selected events (sample with 383 events as mentioned in 7.2.3). The orange nodes are the network metrics, the grey nodes are the meteorological properties, the blue edges represent positive correlations, and the red edges are negative correlations. The thicker the edges, the higher the correlation modulus. Only correlation coefficients above 0.4 or below -0.4 are included in the graph. Every edge connects a network node to a meteorological node, resulting in a bipartite graph.

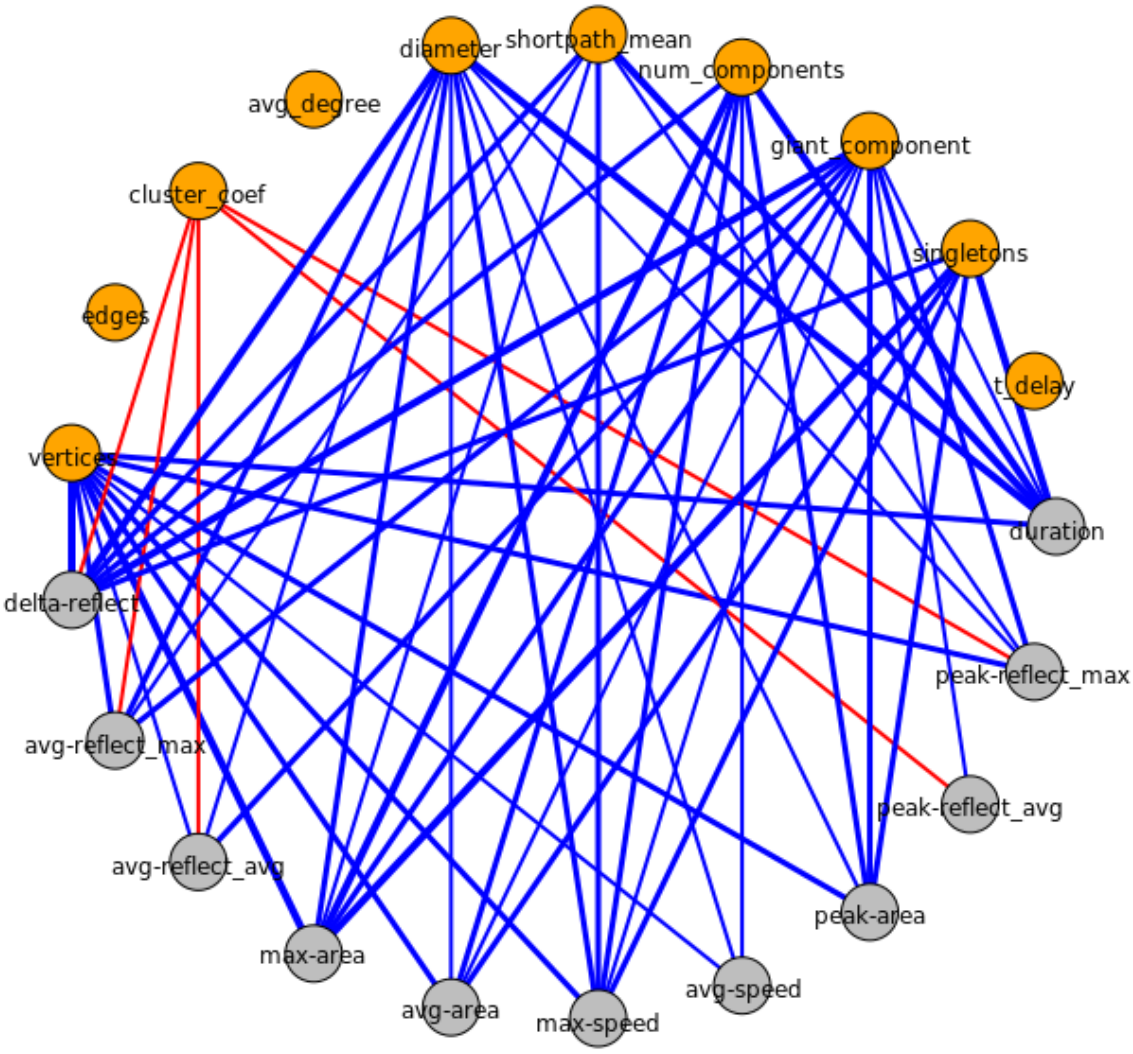
The orange node "t_delay" represents the time delay that maximized the correlations of each network. Unfortunately it did not correlate to any meteorological property. "Vertices", "giant_component", and "diameter" are the nodes with the most significant number of connections, each one with 10. As the event has a larger area, greater duration, or higher speed, the correspondent network tends to spread to follow the event track. Consequently, it results in a higher number of vertices. The result shows that the number of components also tends to increase as well as the size of the giant component. The paths also tend to expand with a more extensive network as the clustering coefficient does not present a positive correlation. Therefore, the diameter metric naturally increases with the event's area, speed, or duration.

We can also notice that, the higher the reflectivity values are ("delta-reflect", "avg-reflect_max", "avg-reflect_avg", "peak-reflect_max"), the greater the network is, as there is a positive correlation with "vertices" property. It shows that events with a wider reflectivity range tend to be those with larger areas or greater duration. For the same reason, this reflectivity variation also positively correlates with the diameter and the average shortest path.

Differently, the clustering coefficient has a negative correlation with the reflectivity

measures. The higher the reflectivity values are, the less clustered the network is. Higher values in reflectivity time series probably become more challenging to have similarities, affecting the network clustering.

Figure 7.4 - Meteo-Network Graph: Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively).



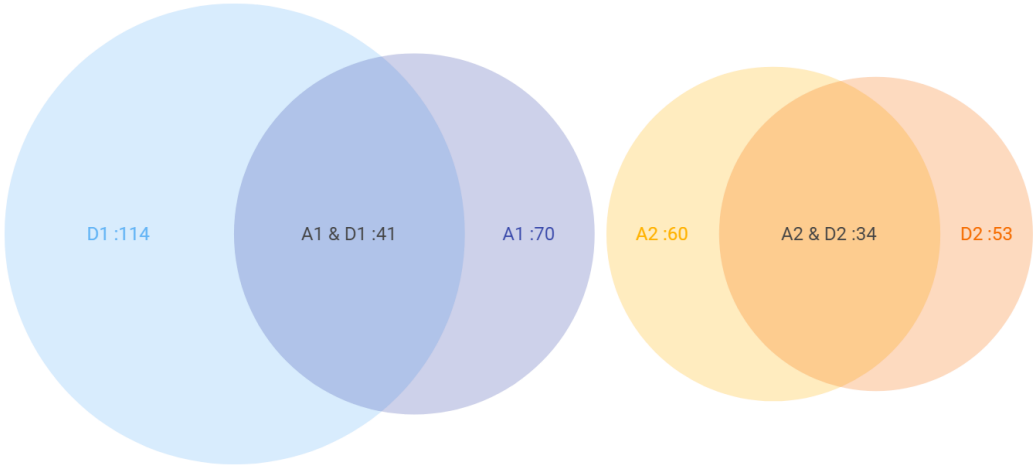
However, the mentioned results are derived from the group of events as a whole, including all kinds of meteorological processes. Intending to analyze these correlations in more specific scenarios, we define four group categories by classifying events by their area size or duration. Table 7.1 describes these groups and Figure 7.5 shows

their intersections. It is possible to notice that short-duration events ($D1$) have a high intersection with small extension ones ($A1$), as well as, most of the long-duration events ($D2$) are included in the group of events with a larger area ($A2$).

Table 7.1 - Groups of Events.

Group	**Filter**	**Number of Events**
D1	Duration $\leq$ 2 hours	114
D2	Duration $>$ 5 hours	53
A1	Area $\leq$ 300 km^2	70
A2	Area $\geq$ 5000 km^2	60

Figure 7.5 - Groups of events and their intersections.

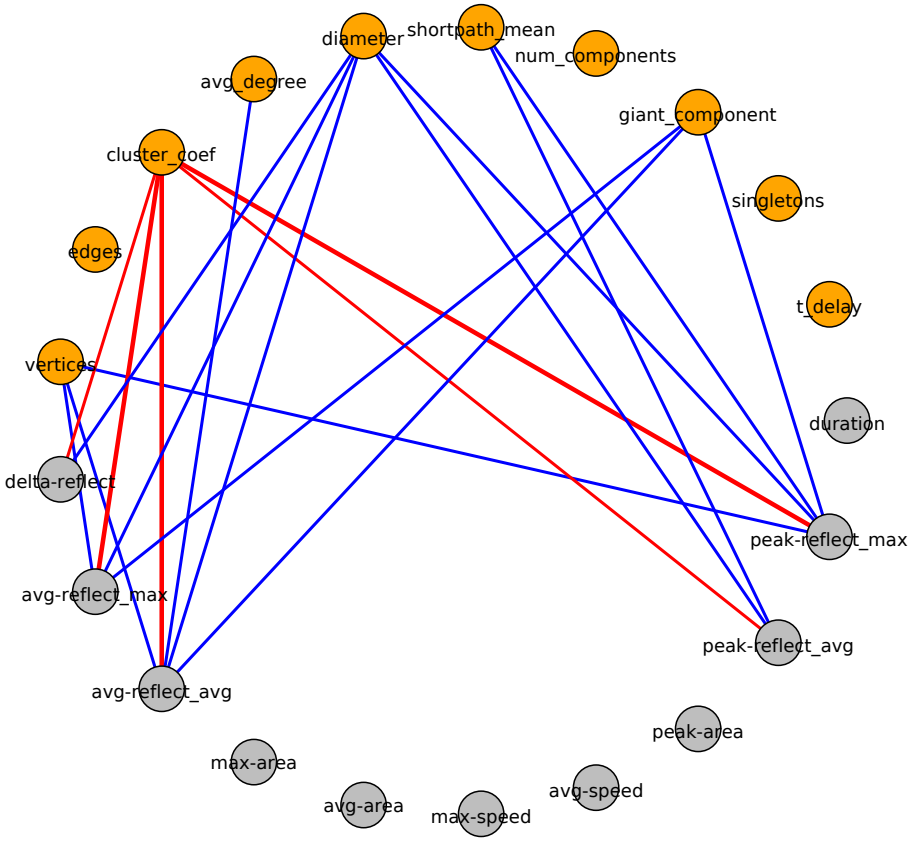


Usually, long-duration events might be associated with weather fronts, which could also reflect in large extensions. On the other hand, short-duration and small area events may be related to local convective systems, which generally present high intensity in a very brief occurrence. As our dataset comprises basically the summer period, convective systems are naturally expected since air humidity is higher in this year's season.

Figure 7.6 shows the graph concerning the D1 group, which concentrates the shortest duration events (2 hours or less). We can observe some particularities when comparing it with the general graph. One of them is the positive correlation between the "avg_degree" and the "avg-reflect_avg". In short events, the average reflectivity appears to increase homogeneously, supporting more connections. The positive

correlations involving area, speed, or duration do not appear for brief events.

Figure 7.6 - Meteo-Network Graph - Group D1 (Short Duration): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively).

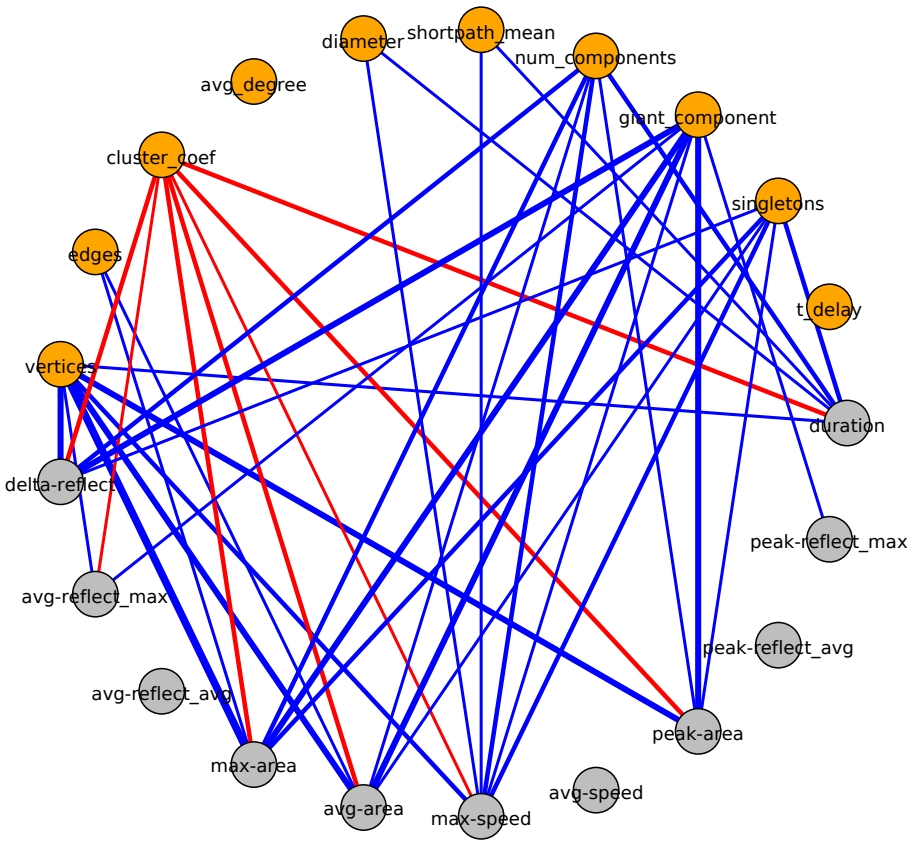


D2 is the group with the longest duration events. Figure 7.7 presents the correspondent graph. In this context of long events, the clustering coefficient has a negative correlation with duration, area ("avg", "max" and "peak"), and "max-speed", which we do not see in the general graph. The diameter and average shortest paths do not correlate with the area size, as the network paths are more related to the duration in this scenario. The positive correlation between duration and "diameter"/"shortpath-mean" corroborates that. On the other hand, the edges positively correlate with maximum and average areas. Similarly, the number of connected components, singletons, and the giant components' size correlate to the area dimension.

Figure 7.8 brings the graph for A1 group, which includes the events with small area (under 300 km^2). Despite having very few edges, we can highlight the correlation

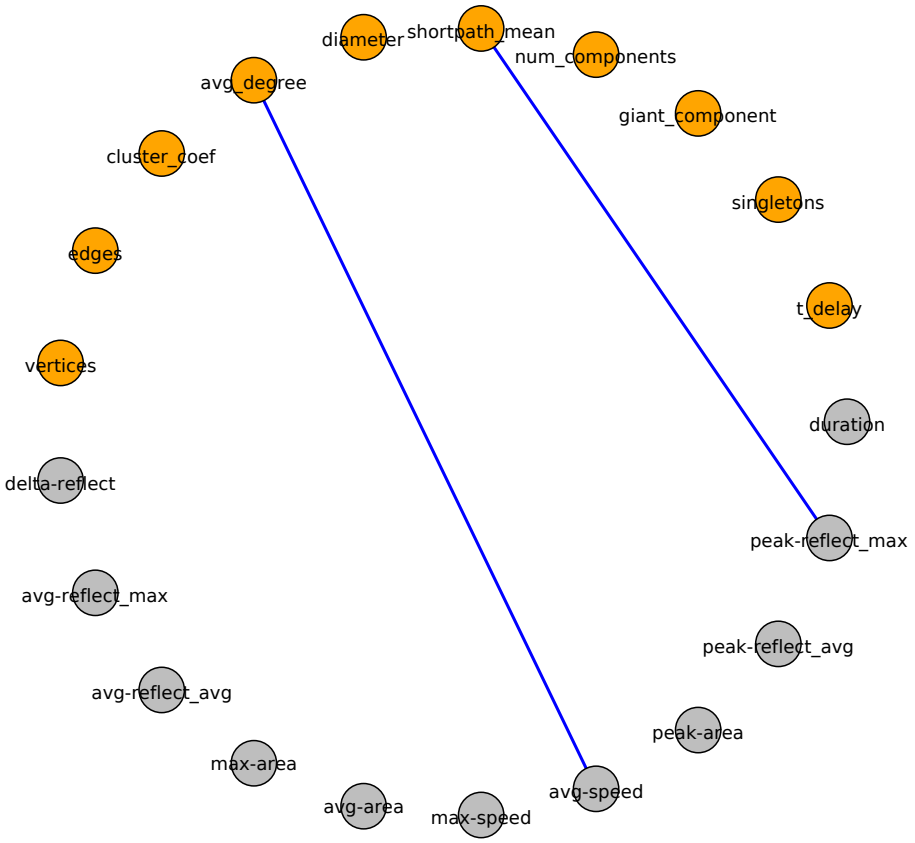
between the average degree and the average speed. As we consider a time delay for the correlations, the speed supports creating more connections. In short extension events, which turn into networks with a few nodes, these connections easily reflect an increase in the average degree.

Figure 7.7 - Meteo-Network Graph - Group D2 (Long Duration): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively).



The last group, A2, includes the events with more extensive areas (above $5000km^2$). In this case, the clustering coefficient has a negative correlation with duration and area ("avg", "max", and "peak"), besides some reflectivity measures. The more extended the network is, the less clustered it is. It is the same behavior we see in the D2 graph. Oppositely, the edges positively correlate with the maximum reflectivity at the event peak. Moreover, the average degree also correlates with the average reflectivity at the event peak. In this group, an increase in the reflectivity values somehow promotes the creation of more connections.

Figure 7.8 - Meteo-Network Graph - Group A1 (Short Extension): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively).



7.4 Final considerations

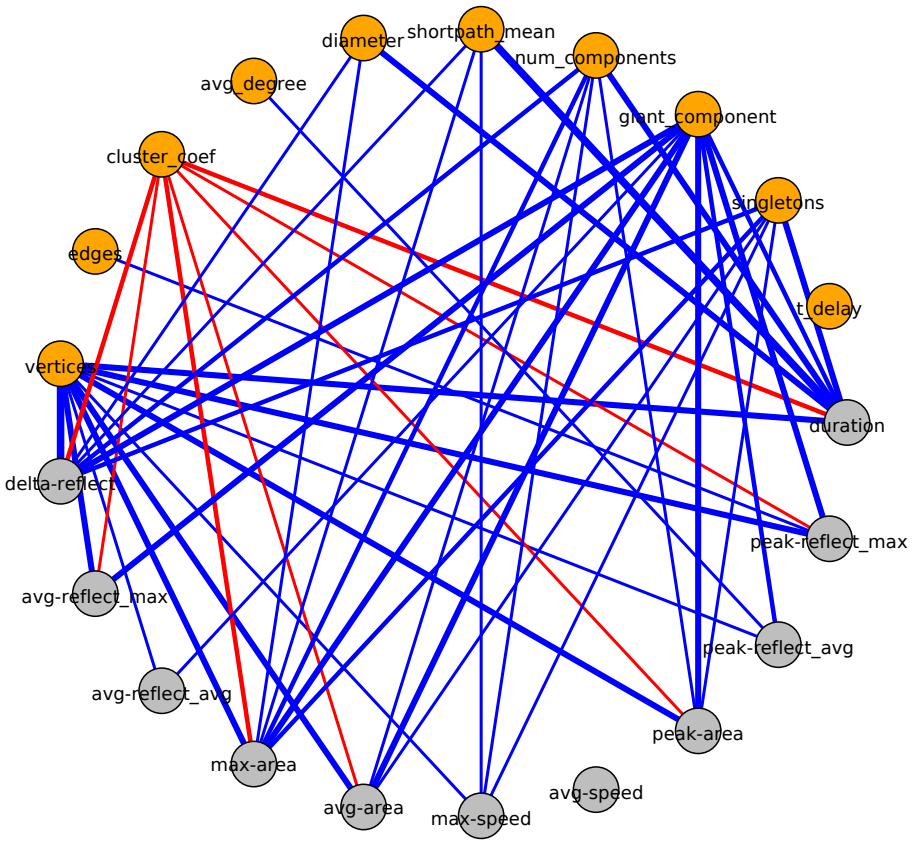
This work presented the analysis of network metrics applied in the weather scope based on a set of precipitation events. We examined the relations between meteorological properties and network metrics. The spatial and temporal components were considered when building up the networks. We identified the relations for the set of events as a whole and groups of events.

Concerning the general ensemble, we could see that a larger area, greater duration, or higher speed tends to extend the correspondent network following the event track. Consequently, metrics such as the number of components and the diameter also tend to increase. The clustering coefficient, otherwise, decreases as the reflectivity values vary.

We also analyzed the meteo-network relations inside specific events groups, classi-

fying them by duration or area size. It was possible to notice some particularities in these scenarios. In short-duration events, the average reflectivity seems to increase homogeneously, supporting more connections and increasing the network's average degree. Concerning long-duration and large-extension events, there is a negative correlation between the clustering coefficient and the event area. In other words, the more extended the network is, the less clustered it is. In short extension events, the speed tends to increase the average degree.

Figure 7.9 - Meteo-Network Graph - Group A2 (Long Extension): Network metrics (orange nodes), Meteorological properties (grey nodes), positive and negative correlations (blue and red edges, respectively).



As a continuation of this work, we plan to analyze the projections of the network and meteorological sets in the bipartite graphs, aiming to go deeper into the correlation analysis. We also intend to embrace other periods in future works and combine different atmospheric variables in multi-layer networks, including forecast data. The expectation is to anticipate extreme events in a very short term.

8 FINAL REMARKS

Extreme weather events can impact society in several ways. The use of complex networks in the weather domain could be a tool to help mitigate these impacts. However, a few works have studied the behavior of networks in such a context. Therefore, this research aimed to answer the scientific question: "What is the behavior of meteorological processes in precipitation networks?" To answer that, we presented three case studies analyzing the behavior of network structures related to precipitation time series. For all of them, weather radar data were employed. The geographical and temporal aspects of the networks were considered when building them up. The first case study analyzes a precipitation network based on a month time series above the Tamanduateí basin. We could verify the spatial dependence of temporal correlations inherent in a precipitation network, noticing a high temporal correlation, especially up to 10 kilometers in geographical distance. We could also explore the relations between topological and geographical distances, observing long topological distances between neighboring nodes and edges connecting short euclidean distances.

The second case study was based on the same dataset from the first case study — the same period and spatial domain. It compared different similarity measures — Mutual Information (MI) and Pearson Correlation (PC) — and criteria – Global-Threshold (GT), Backbone (BB), and Configuration Model (CM) — for building up the networks. Our findings showed that GT and BB criteria produced networks significantly complementary in the geographical space. Moreover, we verified that the combination of MI and BB generated a structure that could be statistically classified as a small-world network.

The last case study described the relations between topological metrics and meteorological properties in a series of precipitation events. These events were tracked at the Metropolitan Area of São Paulo (MASP) from January to March 2019. In a general context, we observed that a larger area, greater duration, or higher speed influenced the network extension, as it tends to follow the event track. As a result, the diameter and the number of components increased. Differently, the clustering coefficient presented a negative correlation with properties related to reflectivity variation. Analyzing more specific scenarios, we could identify some interesting particularities. For example, in the context of short extension events, we noticed that speed tends to increase the number of connections inside the network. Concerning long-duration and large extension events, the area negatively correlates with network

clusterization.

With the presented case studies, this research explored the behavior of network structures in a meteorological context. As a result, we have a basis for future researches in the scope of complex networks applied to anticipate extreme weather events. As our next steps, we intend to employ forecast data, such as an extrapolation from weather radar scans, combined with lightning information from satellite sensors. The expectation is to incorporate these data into a multi-layer geographical network, with the primary goal of promoting the identification of extreme events in the short term. As an additional ideal, we can attempt to classify the events accordingly to the type of meteorological process involved.

REFERENCES

AGARWAL, A.; GUNTU, R.; BANERJEE, A.; GADHAWE, M.; MARWAN, N. A complex network approach to study the extreme precipitation patterns in a river basin. v. 32, 01 2022. 14, 15

AKBAR, S.; SARITHA, S. K. Quantum inspired community detection for analysis of biodiversity change driven by land-use conversion and climate change. **Scientific Reports**, v. 11, n. 1, p. 14332, Jul 2021. ISSN 2045-2322. Available from: <https://doi.org/10.1038/s41598-021-93122-x>. 14

ALBERT, R.; BARABÁSI, A. L. Statistical mechanics of complex networks. **Reviews of Modern Physics**, v. 74, n. 1, p. 47–97, 2002. ISSN 00346861. 1, 11

AMERICAN METEOROLOGY SOCIETY (AMS). **Glossary of meteorology**. Jun 2018. Available from: <http://glossary.ametsoc.org/wiki/Main_Page>. 5

ANDERSEN, T.; SHEPHERD, M. Floods in a changing climate. **Geography Compass**, v. 7, 02 2013. 6

BARABÁSI, A.-L.; PÓSFAI, M. **Network science**. Cambridge: Cambridge University Press, 2016. ISBN 9781107076266 1107076269. Available from: <http://barabasi.com/networksciencebook/>. 1, 6, 7, 9, 11, 23, 29, 33

BOERS, N.; BOOKHAGEN, B.; BARBOSA, H. M.; MARWAN, N.; KURTHS, J.; MARENGO, J. A. Prediction of extreme floods in the eastern Central Andes based on a complex networks approach. **Nature Communications**, v. 5, p. 1–7, 2014. ISSN 20411723. Available from: <http://dx.doi.org/10.1038/ncomms6199>. 1, 2, 14, 15, 29, 43

BOERS, N.; GOSWAMI, B.; RHEINWALT, A.; BOOKHAGEN, B.; HOSKINS, B.; KURTHS, J. Complex networks reveal global pattern of extreme-rainfall teleconnections. **Nature**, v. 566, n. 7744, p. 373–377, 2019. ISSN 14764687. 1, 2, 14, 15, 23, 29

CAVALCANTI, I. F. A.; FERREIRA, N. J.; SILVA, M. G. A. J.; DIAS, M. A. F. d. S. **Tempo e clima no Brasil**. [S.l.]: Oficina de Textos, 2009. ISBN 978-85-86238-92-5. 5, 6

CERON, W.; SANTOS, L. B. L.; DOLIF NETO, G.; QUILES, M. G.; CANDIDO, O. A. Community detection in very high-resolution meteorological networks.

IEEE Geoscience and Remote Sensing Letters, p. 1–4, 2019. ISSN 1545-598X. Available from: <https://ieeexplore.ieee.org/document/8930617/>. 2, 8, 15, 23, 29, 43

COELHO, T. A. S. **Análise geoespacial e mapeamento da densidade de pontos de alagamento em vias públicas do município de São Paulo, entre 2008 e 2013**. PhD Thesis (Geosciences) — Universidade Estadual de Campinas, Campinas/SP - Brazil, 2016. 30

DEPARTAMENTO DE CONTROLE DO ESPAÇO AÉREO (DECEA). **Manual de procedimentos operacionais do radar meteorológico**. [S.l.]: DECEA, 2010. 24 p. 20

DONGES, J. F.; ZOU, Y.; MARWAN, N.; KURTHS, J. Complex networks in climate dynamics. **The European Physical Journal Special Topics**, v. 174, n. 1, p. 157–179, 2009. ISSN 1951-6355. Available from: <http://www.springerlink.com/index/10.1140/epjst/e2009-01098-2>. 1, 13

ENORÉ, D.; COSTA, I.; MACHADO, L. A.; FIGUEIREDO, M.; JORGE, A.; SILVA, D.; SANTOS, D. **Nowcasting: plataforma de previsão imediata do CPTEC/INPE**. 2018. Available from: <http://chuvaproject.cptec.inpe.br/soschuva/pdf/relatorios/relatorio-2019/anexo4.pdf>. 2

FERREIRA, L. N.; FERREIRA, N. C. R.; MACAU, E. E. N.; DONNER, R. V. The effect of time series distance functions on functional climate networks. **The European Physical Journal Special Topics**, v. 230, p. 2973–2998, 2021. 29

IGRAPH. **Igraph - The network analysis package**. 2020. Available from: <https://igraph.org/>. 16

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. **Portal do IBGE**. 2011. Available from: <http://www.ibge.gov.br>. 45

JHA, S. K.; SIVAKUMAR, B. Complex networks for rainfall modeling: spatial connections, temporal scale, and network size. **Journal of Hydrology**, v. 554, p. 482–489, nov 2017. ISSN 00221694. 1, 14, 15

JORGE, A. A. S.; COSTA, I. C.; SANTOS, L. B. L. Geographical Complex Networks applied to describe meteorological data. In: BRAZILIAN SYMPOSIUM ON GEOINFORMATICS, 21., 2020. **Proceedings...** São José dos Campos, Brazil: INPE, 2020. p. 258–263. 29, 43

KRASKOV, A.; STÖGBAUER, H.; GRASSBERGER, P. Estimating mutual information. **Physical Review E**, v. 69, p. 066138, Jun 2004. 8, 29

LIMA, G. N.; RUEDA, V. O. M. The urban growth of the metropolitan area of sao paulo and its impact on the climate. **Weather and Climate Extremes**, v. 21, p. 17–26, 2018. ISSN 2212-0947. 45

LINFORTH, P. **The Digital Artist**. 2020. Available from: <https://pixabay.com/pt/users/TheDigitalArtist-202249/?utm_source=link-attribution&utm_medium=referral&utm_campaign=image&utm_content=3866609>. Access in: 02 Feb. 2020. 1

LOVELACE, R.; ELLISON, R. **stplanr: A package for transport planning**. 2018. Available from: <https://journal.r-project.org/archive/2018/RJ-2018-053/index.html>. 16

MALIK, N.; BOOKHAGEN, B.; MARWAN, N.; KURTHS, J. Analysis of spatial and temporal extreme monsoonal rainfall over South Asia using complex networks. **Climate Dynamics**, v. 39, n. 3, p. 971–987, 2012. ISSN 09307575. 1

MARSHALL, J. S.; LANGILLE, R. C.; PALMER, W. M. Measurement of rainfall by radar. **Journal of Meteorology**, v. 4, p. 186–192, 1947. 20

MENCZER, F.; FORTUNATO, S.; DAVIS, C. A. **A first course in network science**. [S.l.]: Cambridge University Press, 2020. 8, 41

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION (NOAA). **Cloud classification**. NOAA's National Weather Service, Jun 2015. Available from: <https://www.weather.gov/lmk/cloud_classification>. 5

NEO4J. **The fastest path to graph**. 2020. Available from: <https://neo4j.com/>. 17

NETWORKX. **Software for complex networks**. 2019. Available from: <https://networkx.github.io/>. 16

NEWMAN, M. E. J. The structure and function of complex networks. **SIAM Review**, v. 45, n. 2, p. 167–256, 2003. Available from: <https://doi.org/10.1137/S003614450342480>. 9, 32

NUGENT, A.; DECOU, D.; RUSSEL, S. **Atmospheric science**. [s.n.], 2019. Available from: <http://pressbooks-dev.oer.hawaii.edu/atmo/>. 5

OSMNX. **OSMnx 1.1.2 documentation**. 2021. Available from: <https://osmnx.readthedocs.io/en/stable/>. 16

PADGHAM, M. dodgr: An r package for network flow aggregation. **Transport Findings**, 2019. 16

PALUŠ, M.; HARTMAN, D.; HLINKA, J.; VEJMELKA, M. Discerning connectivity from dynamics in climate networks. **Nonlinear Processes in Geophysics**, v. 18, n. 5, p. 751–763, 2011. ISSN 10235809. 1

PEDERSEN, T. L. **Tidygraph**. GitHub, Oct 2019. Available from: <https://github.com/thomasp85/tidygraph>. 16

RAMALHO, D. Rio Tamanduateí - nascente à foz: percepções da paisagem e processos participativos. **Paisagem e Ambiente**, n. 24, p. 99, 2007. ISSN 0104-6098. 24, 30

REDEMET. Redemet, Jun 2015. Available from: <https://www.redemet.aer.mil.br/?i=blog&id=2390>. 20

SANTOS, L. B. L.; CARVALHO, L. M.; SERON, W.; COELHO, F. C.; MACAU, E. E.; QUILES, M. G.; MONTEIRO, A. M. V. How do urban mobility (geo)graph's topological properties fill a map? **Applied Network Science**, v. 4, n. 1, p. 91, Oct 2019. ISSN 2364-8228. Available from: <https://doi.org/10.1007/s41109-019-0211-7>. 8, 31

SANTOS, L. B. L.; CARVALHO, T.; ANDERSON, L. O.; RUDORFF, C. M.; MARCHEZINI, V.; LONDE, L. R.; SAITO, S. M. An rs-gis-based comprehensiveimpact assessment of floods—a case study in madeira river, western brazilian amazon. **IEEE Geoscience and Remote Sensing Letters**, v. 14, n. 9, p. 1614–1617, 2017. 23

SANTOS, L. B. L.; JORGE, A. A. S.; ROSSATO, M.; SANTOS, J. D.; CANDIDO, O. A.; SERON, W.; SANTANA, C. N. de. (geo)graphs - Complex Networks as a shapefile of nodes and a shapefile of edges for different applications. arXiv, v. 321124491, n. November, 2017. Available from: <http://arxiv.org/abs/1711.05879>. 20, 29

SEVTSUK, A.; MEKONNEN, M. Urban network analysis. a new toolbox for arcgis. **Revue Internationale de Géomatique**, v. 22, p. 287–305, 06 2012. 17

SHANNON, C. E. A mathematical theory of communication. **Bell System Technical Journal**, v. 27, n. 3, p. 379–423, 1948. 8

SPNETWORK. **An introduction to the spnetwork package**. 2016. Available from: <https://jsta.github.io/spnetwork/articles/spn.html>. 16

STEINHAEUSER, K.; CHAWLA, N. V.; GANGULY, A. R. An exploration of climate data using complex networks. **ACM SIGKDD Explorations Newsletter**, v. 12, n. 1, p. 25, 2010. ISSN 19310145. 1, 2, 13, 15

STROGATZ, S. H. Exploring complex networks. **Nature**, v. 410, n. 6825, p. 268–276, 2001. ISSN 1476-4687. Available from: <https://doi.org/10.1038/35065725>. 1

SYLVESTER, J. J. Chemistry and algebra. **Nature**, v. 17, p. 284, 1877–8. 6

TSONIS, A. A.; SWANSON, K. L.; ROEBBER, P. J. What do networks have to do with climate? **Bulletin of the American Meteorological Society**, v. 87, n. 5, p. 585–595, 2006. ISSN 00030007. 1, 2, 13, 15, 23, 29, 43

TUTTE, W. T. **Graph theory**. Cambridge, United Kingdom: Cambridge University Press, 2001. (83-12210, v. 21). 7

UBA, D.; GALANTE, R. **Tathu**. GitHub, 2021. Available from: <https://github.com/uba/tathu>. 44

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world' networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998. ISSN 1476-4687. Available from: <https://doi.org/10.1038/30918>. 1

WORLD METEOROLOGICAL ORGANIZATION (WMO). **Climate**. Nov 2019. Available from: <https://public.wmo.int/en/our-mandate/climate>. 5

ZHENG, D.; MHEMBERE, D.; BURNS, R.; VOGELSTEIN, J.; PRIEBE, C. E.; SZALAY, A. S. FlashGraph: processing billion-node graphs on an array of commodity SSDs. In: USENIX CONFERENCE ON FILE AND STORAGE TECHNOLOGIES, 2015. **Proceedings...** Santa Clara/CA, United States, 2015. p. 45–58. ISBN 9781931971201. 17

PUBLICAÇÕES TÉCNICO-CIENTÍFICAS EDITADAS PELO INPE

Teses e Dissertações (TDI)

Teses e Dissertações apresentadas nos Cursos de Pós-Graduação do INPE.

Notas Técnico-Científicas (NTC)

Incluem resultados preliminares de pesquisa, descrição de equipamentos, descrição e ou documentação de programas de computador, descrição de sistemas e experimentos, apresentação de testes, dados, atlas, e documentação de projetos de engenharia.

Propostas e Relatórios de Projetos (PRP)

São propostas de projetos técnico-científicos e relatórios de acompanhamento de projetos, atividades e convênios.

Publicações Seriadas

São os seriados técnico-científicos: boletins, periódicos, anuários e anais de eventos (simpósios e congressos). Constam destas publicações o Internacional Standard Serial Number (ISSN), que é um código único e definitivo para identificação de títulos de seriados.

Pré-publicações (PRE)

Todos os artigos publicados em periódicos, anais e como capítulos de livros.

Manuais Técnicos (MAN)

São publicações de caráter técnico que incluem normas, procedimentos, instruções e orientações.

Relatórios de Pesquisa (RPQ)

Reportam resultados ou progressos de pesquisas tanto de natureza técnica quanto científica, cujo nível seja compatível com o de uma publicação em periódico nacional ou internacional.

Publicações Didáticas (PUD)

Incluem apostilas, notas de aula e manuais didáticos.

Programas de Computador (PDC)

São a seqüência de instruções ou códigos, expressos em uma linguagem de programação compilada ou interpretada, a ser executada por um computador para alcançar um determinado objetivo. Aceitam-se tanto programas fonte quanto os executáveis.