MINISTÉRIO DA CIÊNCIA E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2024/04.02.12.30-TDI

STUDY OF LINKS BETWEEN PEOPLE IN URBAN AREAS BASED ON MOBILITY DATA FOR THE CITY OF SÃO PAULO

Matheus de Moraes Gonçalves Correia

Master's Dissertation of the Graduate Course in Applied Computing, guided by Drs. Leonardo Bacelar Lima Santos, and Vander Luis de Souza Freitas, approved in March 26, 2024.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34T/4B3HSKL>

INPE
São José dos Campos
2024

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 E TECNOLOGIA
INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

sid.inpe.br/mtc-m21d/2024/04.02.12.30-TDI

STUDY OF LINKS BETWEEN PEOPLE IN URBAN AREAS BASED ON MOBILITY DATA FOR THE CITY OF SÃO PAULO

Matheus de Moraes Gonçalves Correia

Master's Dissertation of the Graduate Course in Applied Computing, guided by Drs. Leonardo Bacelar Lima Santos, and Vander Luis de Souza Freitas, approved in March 26, 2024.

URL of the original document:
<http://urlib.net/8JMKD3MGP3W34T/4B3HSKL>

INPE
São José dos Campos
2024

Cataloging in Publication Data

Correia, Matheus de Moraes Gonçalves.

C817s Study of links between people in urban areas based on mobility data for the city of São Paulo / Matheus de Moraes Gonçalves Correia. – São José dos Campos : INPE, 2024.

xx + 52 p. ; (sid.inpe.br/mtc-m21d/2024/04.02.12.30-TDI)

Dissertation (Master in Applied Computing) – Instituto Nacional de Pesquisas Espaciais, São José dos Campos, 2024.

Guiding : Drs. Leonardo Bacelar Lima Santos, and Vander Luis de Souza Freitas.

1. Complex networks. 2. Contact networks. 3. Mobility patterns. 4. Community detection. I.Título.

CDU 004.72(815.6)

MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS

DEFESA FINAL DE DISSERTAÇÃO DE MATHEUS DE MORAES GONÇALVES CORREIA BANCA Nº 033/2024, REG. 355730/2022

No dia 26 de março de 2024, 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, com a exigência de que o trabalho final a ser publicado deverá incorporar as correções sugeridas pela Banca Examinadora, com revisão pelo(s) orientador(es).

Título: "Study of links between people in urban areas based on mobility data for the city of São Paulo"

Membros da Banca:

Dr. Pedro Ribeiro de Andrade Neto – Presidente – INPE
Dr. Leonardo Bacelar Lima Santos – Orientador(a) – INPE
Dr. Vander Luis de Souza Freitas – Orientador(a) – UFOP
Dr. Nandamudi Lankalapalli Vijaykumar – Membro Interno – INPE
Dr. Thadeu Josino Pereira Penna – Membro Externo – UFF

Documento assinado eletronicamente por **Thadeu Josino Pereira Penna (E)**, **Usuário Externo**, em 02/04/2024, às 12:04 (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 02/04/2024, às 12:11 (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 **Pedro Ribeiro de Andrade Neto**, **Tecnologista**, em 02/04/2024, às 12:29 (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 **nandamudi lankalapalli vijaykumar (E)**, **Usuário Externo**, em 02/04/2024, às 13:36 (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 02/04/2024, às 14:20 (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 https://sei.mcti.gov.br/verifica.html, informando o código verificador **11813532** e o código CRC **A09DF3D0**.

Referência: Processo nº 01340.002466/2024-37 SEI nº 11813532

"We each need to find our own inspiration. Sometimes it is not easy".

Hayao Miyazaki

in *"Kiki's Delivery Service"*, 1989

ACKNOWLEDGEMENTS

To my parents, grandparents, and my twin siblings, I thank you for the support you have given me throughout my life. One thing my parents always encouraged in me, my brother, and sister was education, and this is reflected in our upbringing.

To my friends, both those I have kept and those I have found on this journey of the master's degree, I have only gratitude. For moments of relaxation and laughter, but also for being close when I needed it most, thank you all. If I could, I would mention everyone by name, but my heart wouldn't bear it.

To my advisors, Dr. Leonardo Bacelar Lima Santos and Dr. Vander Luis de Souza Freitas, I thank you for all the support, discussions, and nudges. I have grown a lot with your teachings, and I hope to live up to the great instructors you have been to me.

To the academic coordinator, Dr. Rafael Duarte Coelho dos Santos, and the staff of the Applied Computing Postgraduate Program at INPE, who have always been ready to clarify my doubts and answer my numerous emails.

Finally, I would like to thank CAPES scholarship 88887.668467/2022-00 and the CNPq project 446053/2023-6 for all the support. Ending with one last thank you to the Brazilian Ministry of Science, Technology, and Innovation, and the Brazilian Space Agency.

ABSTRACT

Complex networks have made significant contributions to our understanding of the properties and dynamics of real-world networks. They have been instrumental in understanding mobility patterns in large urban centers, as well as studying the transmission of infectious diseases, which has become increasingly important since the COVID-19 pandemic in 2020. This study delves into the contact networks derived from urban mobility data in São Paulo, Brazil. Leveraging an Origin-Destination (OD) dataset, we investigate the structural dynamics of contact networks and their correlation with residential components over various temporal resolutions. The study employs the Louvain algorithm for community detection, providing insights into the regional structuring of the city. The analysis begins by scrutinizing the contact network's behavior throughout the day, unveiling distinct temporal patterns and highlighting pivotal moments of heightened activity. It then explores the impact of varying minimum contact durations (mcd) on network properties, revealing the transformation of the network's connectivity and the emergence of critical timeframes for urban movement. Additionally, a detailed analysis of network metrics, including degree, clustering coefficient, and strength, offers a comprehensive understanding of the network's topological features. These metrics shed light on the connectivity patterns, local clustering behavior, and the influence of contact strength, providing a nuanced perspective on the intricate dynamics of urban contact networks. The implications of this research extend beyond understanding urban mobility patterns, offering potential applications in epidemiology, disaster prevention, and urban planning. By decoding the complex interplay of contact networks and unveiling key network metrics, this study contributes to unraveling the behavior of urban dynamics, providing a foundation for further investigations and practical applications.

Keywords: Complex Networks. Contact Networks. Mobility Patterns. Community Detection.

ESTUDO DE VÍNCULOS ENTRE PESSOAS EM ÁREAS URBANAS COM BASE EM DADOS DE MOBILIDADE PARA A CIDADE DE SÃO PAULO

RESUMO

Redes complexas fizeram contribuições significativas para a nossa compreensão das propriedades e dinâmicas das redes do mundo real. Têm sido fundamentais para a compreensão dos padrões de mobilidade nos grandes centros urbanos, bem como para o estudo da transmissão de doenças infecciosas, que se tornou cada vez mais importante desde a pandemia da COVID-19 em 2020. Este estudo investiga as redes de contato derivadas de dados de mobilidade urbana em São Paulo, Brasil. Aproveitando um conjunto de dados Origem-Destino (OD), investigamos a dinâmica estrutural das redes de contato e sua correlação com componentes residenciais em várias resoluções temporais. O estudo emprega o algoritmo Louvain para detecção de comunidades, fornecendo insights sobre a estruturação regional da cidade. A análise começa examinando o comportamento da rede de contatos ao longo do dia, revelando padrões temporais distintos e destacando momentos cruciais de maior atividade. Em seguida, explora o impacto das diferentes durações mínimas de contato (mcd) nas propriedades da rede, revelando a transformação da conectividade da rede e o surgimento de prazos críticos para o movimento urbano. Além disso, uma análise detalhada das métricas da rede, incluindo grau, coeficiente de aglomeração e força, oferece uma compreensão abrangente dos recursos topológicos da rede. Estas métricas esclarecem os padrões de conectividade, o comportamento dos agrupamentos locais e a influência da força de contacto, proporcionando uma perspectiva diferenciada sobre a intricada dinâmica das redes de contacto urbanas. As implicações desta pesquisa vão além da compreensão dos padrões de mobilidade urbana, oferecendo aplicações potenciais em epidemiologia, prevenção de desastres e planejamento urbano. Ao descodificar a complexa interação das redes de contacto e ao revelar as principais métricas de rede, este estudo contribui para desvendar comportamentos da dinâmica urbana, fornecendo uma base para futuras investigações e aplicações práticas.

Palavras-chave: Redes Complexas. Redes de Contato. Padrão de Mobilidade. Detecção de Comunidade.

LIST OF FIGURES

Page

2.1 Graphical representation of graph. 5
2.2 Graphical representation of a disconnected graph. 6
2.3 Graphical representation of each step for the Louvain method. The first phase is represented by "VM" (Vertex Mover), and the second phase by "Agg." (Aggregation). In this illustration, the algorithm converges after two passes, uncovering an optimal partition made of 2 communities and a value of $Q = 0.45$. 11

3.1 Division of the sub-regions of the OD Survey. 15
3.2 Division of zones and municipalities for each of the sub-regions. 16
3.3 Division of São Paulo's regions into sub-prefectures, districts, and zones. 17
3.4 Regions of São Paulo. 18
3.5 Flowcharts of all processes divided into 3 main blocks, representing respectively data preprocessing, contact network construction, and output. 21

4.1 Number of people per residential zone. 24
4.2 Population density per residential zone. 25
4.3 Number of trips made at each hour of the day, starting from 00:00 and ending at 23:00. 26
4.4 Left: Frequency of the quantity of trips made; Right: Frequency of the number of visited zones. Both graphs represent a day's time span. 26
4.5 The graph displays the number of connected components (NCC) and the largest connected component (LCC) for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. 27
4.6 This figure shows the increase in NCC as the mcd increases. Blue nodes are still connected to the network while red are isolated. 29
4.7 The graph displays the average degree for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the average degree $\langle k \rangle$ of a graph generated with the corresponding threshold value. 30
4.8 The graph displays the density for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the density D of a graph generated with the corresponding threshold value. . 31

4.9 The graph displays the average strength for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the average strength $\langle s \rangle$ of a graph generated with the corresponding threshold value. 32
4.10 The graph displays the average strength for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the average clustering $\langle c \rangle$ of a graph generated with the corresponding threshold value. 33
4.11 In this figure, we have 24 graphs, each one representing the degree of the vertices of the network. Considering that each point represents a vertex, the y-axis contains the number of vertices with the given degree, represented by the x-axis. 34
4.12 In this figure, we have 24 graphs, each one representing the strenght of the vertices of the network. Considering that each point represents a vertex, the y-axis contains the number of vertices with the given strenght, represented by the x-axis. 35
4.13 In this figure, we have 24 graphs, each one representing the clustering of the vertices of the network. Considering that each point represents a vertex, the y-axis contains the number of vertices with the given clustering, represented by the x-axis. 36
4.14 The strength of each vertex varies with the *mcd*. 38
4.15 The outer strength of each vertex varies with the *mcd*. 39
4.16 Each point represents a node, and its color corresponds to each of the communities presented. 41
4.17 Dominance of each zone. 42
4.18 Histogram of dominance values. 43
4.19 Community division by zones. 44

LIST OF TABLES

Page

3.1 Dataset attributes 19
3.2 Profile of the first 5 rows of the dataset of the city of Sao Paulo. 20

4.1 The most common activity performed and primary mode of transportation. . 32
4.2 Comparison of Zone Distributions among Regions and Communities in São Paulo. 45

LIST OF ABBREVIATIONS

CN	–	Complex Networks
OD	–	Origin-Destination
NCC	–	Number of connected components
LCC	–	Largest connected component

CONTENTS

Page

1 INTRODUCTION 1
1.1 Motivation 2
1.2 Objective 2
1.3 Contributions 2
1.4 Organization of the document 3

2 LITERATURE OVERVIEW 5
2.1 Graph theory 5
2.1.1 Connectedness 6
2.1.2 Network density 6
2.1.3 Centralities 7
2.1.4 Clustering coefficient 8
2.1.5 Community detection 9
2.1.6 Fast community unfolding 9
2.1.7 Dominance coefficient 11
2.2 Related works 12

3 MATERIALS AND METHODS 15
3.1 Study area 15
3.2 Origin-Destination data 19
3.3 Network construction 20

4 RESULTS 23
4.1 Data analysis and setting up the contact network 23
4.2 Network metrics 27
4.2.1 Ordinary and outer strength 37
4.3 Community detection 40

5 CONCLUSIONS 47

REFERENCES 49

1 INTRODUCTION

The study of complex networks has made significant contributions to our understanding of the properties and dynamics of real-world networks, from the internet and social networks to biological systems (ALBERT; BARABÁSI, 2002; WATTS; STROGATZ, 1998; COSTA et al., 2011). This field has helped uncover fundamental principles governing the behavior of complex systems. It has the potential to shed light on many pressing issues, such as the spread of diseases, the functioning of economies, and the behavior of social groups, among others.

The emergence of the SARS-CoV-2 virus in early 2020 has resulted in a devastating pandemic that has rapidly spread worldwide, with over 38 million people infected and over 709,000 deaths recorded in Brazil since the first case was reported in São Paulo on February 25, 2020 (COTA et al., 2020). Computational models have become increasingly crucial to deal with the complex problems presented by the pandemic. Mathematical models focused on epidemiology have gained traction as valuable tools for forecasting and highlighting the importance of actions to reduce the number of cases (ALLEN et al., 2008; PINTO et al., 2020). One promising approach in this area is the use of mobile network modeling based on data from the region under study (FREITAS et al., 2020b; BONA et al., 2016).

A mobility network, which consists of locations connected by the flow of people, is essential for understanding virus transmission on a large scale, particularly in a vast country like Brazil (FREITAS et al., 2020a; YILDIRIMOGLU; KIM, 2018; LAMOSA et al., 2021).

From the mobility network, a network of contacts can be built (GONZALEZ et al., 2008; BANSAL et al., 2010), treating individuals as nodes and the contact time between them as edges in a graph. Among other diseases, the transmission of COVID-19 (YANG et al., 2021; BANSAL et al., 2010) from an infected to a healthy person happens through contact, making contact networks fundamental to providing insights into the complex systems of many individuals.

1.1 Motivation

Before delving into the study of disease transmission, it is crucial to understand the patterns of urban mobility and how individuals are interconnected.

In the metropolitan region of São Paulo, a dataset is compiled every ten years by the São Paulo Metropolitan Company (Metrô) in collaboration with the Government of the State of São Paulo. This research initiative began in 1977, and the most recent available data was released in 2017 on the Transparency Portal of Metrô (OD SURVEY, 2017).

This analysis could offer insights into the transmission dynamics of viruses in specific regions and guide the development of targeted interventions to control their spread. Additionally, it could assist in identifying communities at higher risk and allocating resources more effectively to contain epidemics.

1.2 Objective

This study aims to investigate and analyze the structure within a contact network derived from urban mobility data in the city of São Paulo. The research aims to apply community detection algorithms to identify significant clustering patterns, understand the geographical distribution of these communities in relation to city characteristics, and explore potential correlations with factors such as population density, public transportation infrastructure, and mobility patterns.

In this context, the scientific questions to be answered are:

- "What is the relationship between people based on the places they visit?"
- "How does the contact network behave throughout the day? What are the topological properties of the network?"
- "Does the community structure formed follow residential components?"

1.3 Contributions

This thesis can contribute to the understanding of the contact network structure in urban environments, especially for large metropolises like São Paulo. Some contributions include the identification of mobility patterns. By analyzing the contact network structure, patterns of urban mobility can be identified, highlighting the main areas of interaction and movement in the city.

The application of community detection algorithms reveals significant clusters in the network, assisting in understanding how people interact in different regions of the city.

This knowledge of mobility patterns applies not only in epidemiology, as mentioned, but also in areas such as disaster prevention (TOMÁS et al., 2022; SANTOS et al., 2019). This topic aligns with one of the strategic objectives of the National Institute for Space Research (INPE) between 2022 and 2026, based on the goal of promoting the use and dissemination of images, technologies, and space services for disaster management (INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS, 2022).

As part of the development of this dissertation, an abstract (CORREIA et al., 2022) and a full paper (CORREIA et al., 2023) were published at the National Congress of Applied and Computational Mathematics (CNMAC).

1.4 Organization of the document

The thesis is organized according to the following chapters. In Chapter 2, the literature overview provides an in-depth analysis of the theory used to calculate network metrics and community detection, emphasizing the importance of our study.

Chapter 3 introduces the specific area of the study, along with the data sources and methodology used to construct the network. It includes a detailed description of the data cleaning and preprocessing steps, as well as the criteria used to define the network edges and nodes.

In Chapter 4, the results obtained are presented, along with the final considerations. Finally, Chapter 5 provides a conclusion and outlines future works from this thesis.

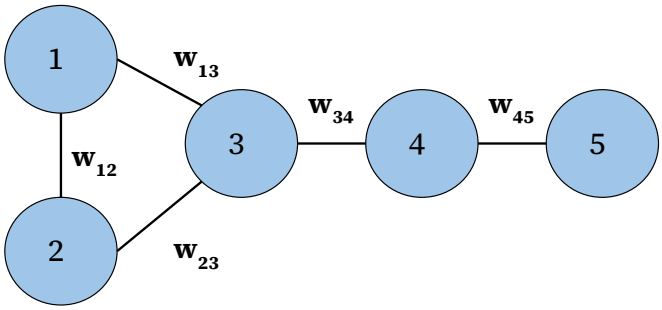
2 LITERATURE OVERVIEW

Graph theory has emerged as a powerful tool in the computational analysis of mobility data. In our study, we utilized this framework to construct a graph where each vertex represents an individual, and the edges represent the relationships between them. The weights of these edges vary depending on the duration of time that the individuals spent in contact with each other. This approach allows us to model the mobility patterns of individuals and provides a way to study the network structure of the resulting graph.

2.1 Graph theory

A network is formally defined as a graph $G(V, E)$, where V represents a non-empty set of vertices (nodes), and E represents a set of non-ordered pairs of vertices, which correspond to the edges (links) of G (BOAVENTURA NETTO, P. O.; JURKIEWICZ, 2017). The adjacency matrix, $A = a_{ij}$ for $i, j = 1, ..., N$, is used to represent the links between vertices in the graph. The value of a_{ij} is 1 if there is an edge between vertices i and j, and 0 otherwise, where $N = |V|$ and $L = |E|$ is the total number of nodes and edges in the graph, respectively. Edge weights are assigned to the edges in a matrix $W = w_{ij}$ for $i, j = 1, ..., N$, where w_{ij} represents the weight of the edge connecting vertices i and j. In the case of our study, the weight is the duration of contact between individuals, and the network is considered undirected, meaning the connections between nodes have no specified direction. An example can be visualized in the graph depicted in Figure 2.1, where vertices are numbered from 1 to 5, with w_{ij} indicating the weight of connections between such vertices.

Figure 2.1 - Graphical representation of graph.



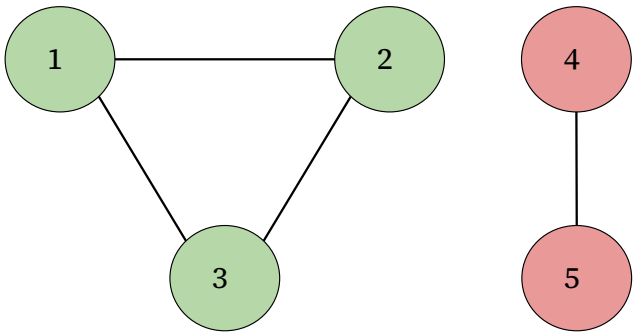
Notably, the adjacency and weight matrices are constructed with a zero value in their main diagonal, which disallows self-loops (i.e., $a_{11}, a_{22}, ..., a_{NN} = 0$ and $w_{11}, w_{22}, ..., w_{NN} = 0$).

2.1.1 Connectedness

In an undirected network, two different nodes, i and j, are connected if there is a path between them. If that path does not exist, they are disconnected. As for the network, it is connected if all its vertices are connected; otherwise, it will be disconnected and have more than one connected component, where each component is a subset of the network nodes (BARABáSI; PóSFAI, 2016).

Knowing this, it is possible to calculate the number of connected components (NCC) that exist in the network and also which would be the largest connected component (LCC) by checking the number of nodes that compose it. As an example, Figure 2.2 illustrates a graph featuring two connected components (NCC=2), with its largest connected component comprising 3 vertices (LCC=3).

Figure 2.2 - Graphical representation of a disconnected graph.



2.1.2 Network density

In graph theory, the density of an undirected network can be defined as the ratio between the number of edges in the graph, L, and the number of possible total edges, which is $N(N-1)/2$, in an undirected network, so that N is the total number of nodes in the network (BARABáSI; PóSFAI, 2016). Mathematically, this can be expressed as:

$$D = \frac{2L}{N(N-1)}. \tag{2.1}$$

A density value of 1 indicates that the network is complete, where an edge connects every pair of nodes, while a value of 0 indicates an empty network with no edges. In our case, we used this measure to analyze the degree of connectivity of the contact network.

2.1.3 Centralities

In an undirected network, the degree of a node k_i represents the number of links that node i has with other nodes. As there are no differences between incoming and outgoing edges in an undirected network, k_i can be interpreted as the number of neighbors that node i has. The degree k_i for node i can be calculated as the sum of the elements in the i-th row (or column, since the matrix is symmetric) of the adjacency matrix A, that is:

$$k_i = \sum_{j=1}^{N} a_{ij}, \tag{2.2}$$

where N is the total number of nodes in the network (BARABáSI; PóSFAI, 2016).

To calculate the average degree of the graph, denoted by $\langle k \rangle$, we can use the formula:

$$\langle k \rangle = \frac{2L}{N}, \tag{2.3}$$

which gives an idea of the connectivity of the network as a whole. A higher value of $\langle k \rangle$ indicates a denser network with more connections, while a lower value indicates a sparser network with fewer connections.

The vertex strength of a node is defined as the sum of the weights of all its connections. In a network where all connections have a weight equal to 1, the vertex strength is simply the degree of the node. However, in some real-world networks, connections may have different weights, representing different levels of importance or influence.

To calculate the vertex strength of a node, we first need to assign a weight to each connection. We then sum the weights of all connections to the node to obtain its

vertex strength (BARRAT et al., 2004). Mathematically,

$$s_i = \sum_{j=1}^{N} a_{ij} w_{ij}. \tag{2.4}$$

The average vertex strength of a network is simply the average of the vertex strengths of all nodes in the network. It represents the typical level of importance or influence of nodes in the network. The average vertex strength of a network is given by:

$$\langle s \rangle = \frac{1}{N} \sum_{i=1}^{N} s_i. \tag{2.5}$$

2.1.4 Clustering coefficient

The clustering coefficient, denoted by C, is a metric commonly used in network analysis to measure the tendency of nodes in a network to form clusters or tightly interconnected groups. It provides insight into the local structure of a network by quantifying the extent to which a node's neighbors are interconnected. Specifically, for a node i with degree k_i, the clustering coefficient is defined as the ratio of the number of links between neighbors of node i to the total number of possible links between them (WATTS; STROGATZ, 1998).

It should be noted that for nodes with degree $k_i = 0$ or $k_i = 1$, the clustering coefficient is undefined since there are no neighbors to form connections between. Thus, by convention, $C_i = 0$ for such nodes (BARABáSI; PóSFAI, 2016). For this study, a weighted clustering coefficient, C^w, was calculated, which takes into account the weights of the edges connected to each vertex (BARRAT et al., 2004):

$$C_i^w = \frac{1}{s_i(k_i - 1)} \sum_{j,h} \frac{(w_{ij} + w_{ih})}{2} a_{ij} a_{ih} a_{jh}, \tag{2.6}$$

in which s_i is the strength of node i.

To obtain the average clustering coefficient of a graph, denoted as $\langle C \rangle$, one needs to divide the sum of the weighted clustering coefficients C_i^w for all nodes i in the network by the total number of nodes N:

$$\langle C \rangle = \frac{1}{N} \sum_{i} C_i^w. \tag{2.7}$$

2.1.5 Community detection

A community is a group of nodes that have a higher probability of connecting rather than to nodes outside of their communities. Keeping this concept in mind, Barabási and Pósfai (2016) proposed four hypotheses for community detection algorithms.

The Fundamental Hypothesis posits that communities are inherent structures encoded within a network's adjacency matrix, A_{ij}. These communities represent underlying truths awaiting discovery through the application of appropriate algorithms.

According to the Connectedness and Density Hypothesis, a community is characterized as a locally dense connected subgraph. It means that nodes within a community exhibit a higher density of connections among themselves compared to nodes outside the community.

The Random Hypothesis asserts that randomly wired networks lack discernible communities; in contrast to networks with meaningful structures, a random network needs a clear community organization.

Finally, the Maximal Modularity Hypothesis suggests that the most accurate representation of a network's community structure is achieved by partitioning the network into communities that maximize modularity.

Modularity (Q) is defined as the quality of the network partition into subgraphs. In other words, it measures the density of links within the community compared to links connecting different communities (NEWMAN, 2004a).

There are various methods for conducting this analysis (FORTUNATO, 2010), with the Girvan and Newman method (NEWMAN; GIRVAN, 2004) being the most popular. However, for large networks, such as the mobility network of a metropolis like São Paulo, a low-complexity method is necessary due to the vast amount of information in this network. Therefore, the Louvain algorithm was used to detect communities (BLONDEL et al., 2008) due to its low computational complexity, $O(NlogN)$, especially considering that the network under study is extensive and composed of thousands of nodes.

2.1.6 Fast community unfolding

Fast community unfolding, also known as the "Louvain Method", is a heuristic method for greedy modularity optimization, a direct use of the aforementioned

fourth hypothesis. In other words, it is a method that seeks partitions (communities) with optimal modularity. The following equation can represent modularity:

$$Q = \frac{1}{2m} \sum_{i,j} [w_{ij} - \frac{t_i t_j}{2m}] \delta(c_i, c_j), \tag{2.8}$$

where $t_i = \sum_j w_{ij}$ is the sum of weights of edges connected to vertex i, and c_i is the community to which vertex i has been assigned. Similarly, t_j and c_j refer to vertex j. The function $\delta(c_i, c_j)$ equals 1 if both vertices are in the same community and 0 otherwise. Finally, $m = \frac{1}{2} \sum_{i,j} w_{ij}$ (NEWMAN, 2004b).

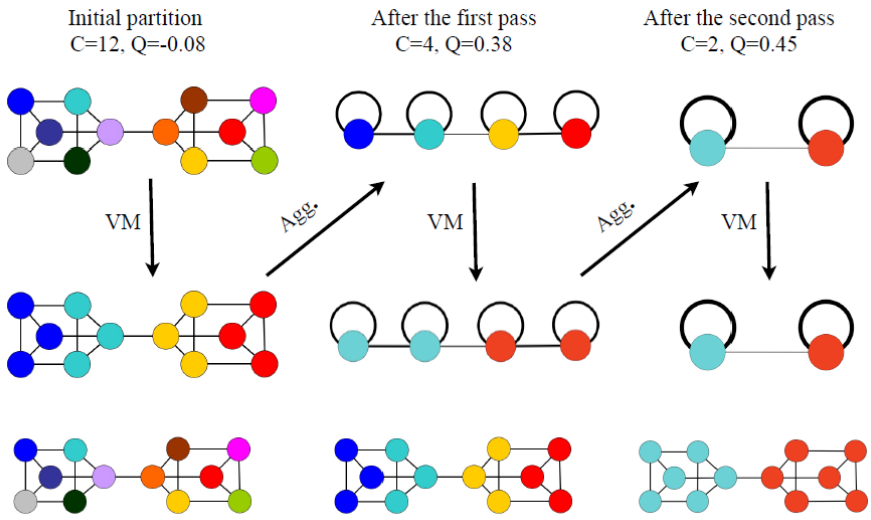
As the method focuses on seeking the best modularity, it operates in a multi-resolution scheme, thus being highly scalable, with a complexity of O(L), where L is the number of links in the graph. It operates by following these steps:

- Assign every node to its community and calculate the modularity for the graph.
- Phase 1: Vertex Mover (VM)
  - Traverse all vertices and check for an increase in modularity by moving the node from its community to a neighbor's community;
  - Place the node in the community that maximizes modularity;
  - Repeat until there are no more gains in modularity.
- Phase 2: Aggregation (Agg.)
  - Nodes from the same communities are merged into a "super-node";
  - The weights of the edges are summed.
- Repeat the previous phases until there is no further increase in modularity.

This algorithm results in optimal graph partitioning in terms of modularity. The method is relatively fast, moving only vertices between neighboring communities. Community exploration is initially performed locally and then over long distances as vertices aggregate, as seen in Figure 2.3.

Additionally, it avoids the problems faced by other methods where the calculation of modularity has some limitations in identifying communities up to a particular scale

Figure 2.3 - Graphical representation of each step for the Louvain method. The first phase is represented by "VM" (Vertex Mover), and the second phase by "Agg." (Aggregation). In this illustration, the algorithm converges after two passes, uncovering an optimal partition made of 2 communities and a value of $Q = 0.45$.



SOURCE: Blondel et al. (2023).

because it depends on the resolution parameter (r) chosen by the user, which can alter the size of the communities. Reducing the resolution reveals smaller clusters, leading to a greater number of them. Conversely, increasing the resolution identifies clusters that include more data points.

2.1.7 Dominance coefficient

Community detection aims to find patterns, and it is expected to uncover relationships between each of these subgraphs. The concept of a "dominance coefficient" refers to how influential or prominent specific communities are in a network. The more significant number of nodes can analyze this from a specific community in a region or zone. Blondel et al. (2008) provides an example by examining the percentage of speakers of the language, French or Dutch, in a studied community.

In our study, the dominance measure on a scale of 0 to 1 indicates the degree of presence or influence of a community in a specific zone. A value closer to 1 suggests the dominance of the community in that zone, while a value closer to 0 indicates a less prominent presence.

This metric is valuable for assessing the influence of a community in different geographic regions or parts of a network. It provides insights into how significant the contribution of a specific community is in terms of connections or interactions in a particular zone.

2.2 Related works

To underscore the significance of incorporating complex networks into the study of mobility, particularly in the context of epidemic diseases such as COVID-19, an extensive search was conducted on Google Scholar using the search query *("complex networks" or "graph") and ("mobility" or "contact") or ("connection" or "interaction")*. By encompassing relevant topics in our research, we aimed to ensure the currency of our findings with ongoing trends.

Among all the resulting articles, the first fifty were selected, and all of them are centered on Complex Networks (CN). Fifteen articles (30%) were categorized under the utilization of CN to investigate disease transmission, specifically in the context of epidemic diseases. Moreover, five out of those fifteen articles (30%) were explicitly related to COVID-19, emphasizing the relevance and timeliness of our research topic.

As exemplified by Pastor-Satorras et al. (2015) and Liu et al. (2013), the recognition of the expanding field of epidemic modeling in networks and the potential for cross-disciplinary application has led to improvements in the study of epidemics. Subsequently, the application of this field has broadened to encompass a diverse range of domains.

In the aftermath of the devastating global impact of the coronavirus pandemic in recent years, numerous studies, including those by Goel et al. (2021) and Hâncean et al. (2020), have emerged, utilizing mobility networks to investigate disease transmission. In addition to these studies, akin to our work focused on Brazil, studies such as Freitas et al. (2020b) and Lamosa et al. (2021) conducted research using mobility networks, with nodes representing places and edges representing the flow between them.

Works such as Pechlivanoglou et al. (2022) and Hartnett et al. (2021) employ mobility data to construct contact networks for the study of coronavirus transmission. In the context of this ongoing pandemic, research in this area continues to be critically important for understanding virus spread and developing effective mitigation strategies.

Complex networks have gained substantial traction in this domain, and our work, incorporating mobility data from the city of São Paulo, holds the potential to enrich this steadily growing field further.

Additionally, when delving into the study of São Paulo, a more comprehensive search was conducted. Utilizing a new search query for the search *("complex networks" OR graphs) AND (mobility OR contact OR connection OR interaction) AND ("São Paulo" OR "Sao Paulo")*, the objective was to review all pertinent literature and confirm the originality of this thesis. In contrast to the former, this study was conducted using the Federated Academic Community (CAFe) access infrastructure at the Coordination for the Improvement of Higher Education Personnel's (CAPES) Periodicals Portal.

In total, 311 article results were obtained, with only one of them utilizing the same data from the 2017 Origin-Destination Survey. Martins et al. (2021), employing the identical dataset as ours, sought to identify and analyze mobility patterns for various scenarios, such as peak travel times, social factors, and transportation modes, unlike our study, where we applied such data to construct a contact network and analyze its topological metrics, including community detection.

By leveraging the interconnectivity of complex networks and the intricacies of epidemic models, a more profound understanding of how diseases spread within populations can be gained. This approach has the potential not only to enhance our comprehension of ongoing epidemics but also to contribute to the development of strategies to combat future outbreaks. Therefore, the study of complex networks in epidemic modeling represents a promising avenue for advancing our knowledge of infectious diseases and improving public health outcomes.

3 MATERIALS AND METHODS

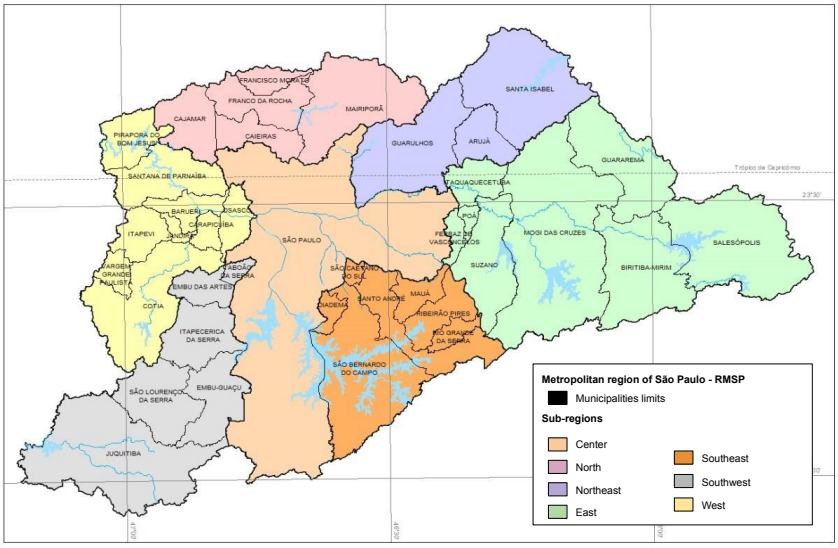
3.1 Study area

To procure the requisite data for our study, we relied on the 2017 Origin-Destination (OD) Survey conducted by the Secretary of Metropolitan Transport of the Government of the State of São Paulo in collaboration with the Metro company (OD SURVEY, 2017). This comprehensive survey covered the entire metropolitan region of São Paulo, furnishing us with a robust dataset for our analysis.

This survey was conducted in various households where residents were asked about various themes, for example, the means of transportation, places, and times they had visited on the previous business day. They interviewed about 0.3% of the population of the city of São Paulo in 2017.

For our study, we focused on the sub-region (Center) of the OD research, depicted in Figure 3.1. This sub-region precisely corresponds to the city of São Paulo, while other cities in the metropolitan region mainly function as distinct sub-regions.

Figure 3.1 - Division of the sub-regions of the OD Survey.

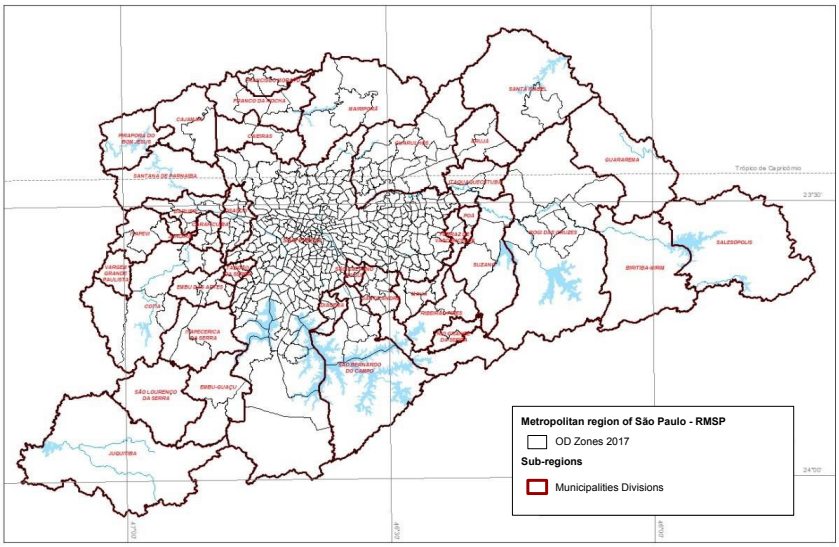


SOURCE: OD SURVEY (2017).

This particular sub-region is comprised of 342 OD zones, defined based on urban and socioeconomic homogeneity, as well as other technical criteria. These zones represent

the minor geographic units, ensuring the statistical representativeness of the data, as depicted in Figure 3.2.

Figure 3.2 - Division of zones and municipalities for each of the sub-regions.



SOURCE: OD SURVEY (2017).

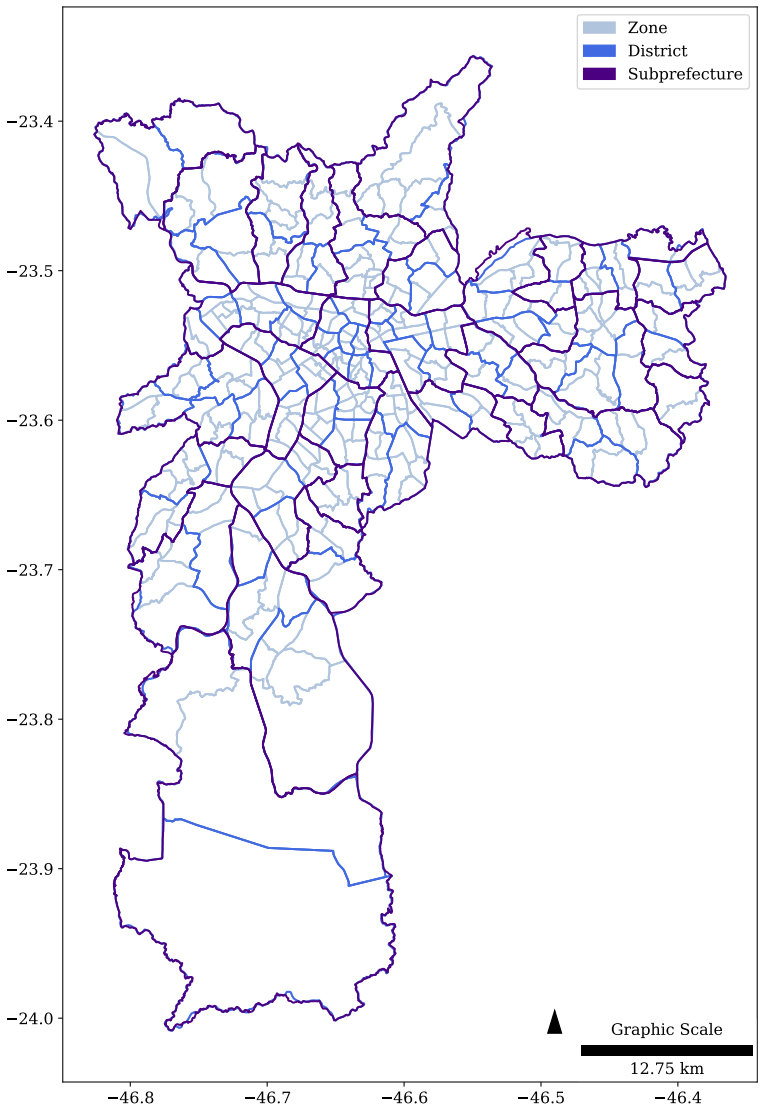
As a complement to the 2017 Origin-Destination Survey, additional support was derived from data provided by the Digital Map of the City of São Paulo (GEOSAMPA, 2017). Developed by the São Paulo City Hall in collaboration with the Municipal Department of Urbanism and Licensing, this map facilitated the acquisition of supplementary data for our thesis.

With this data, we could utilize the city's division into larger geographical units such as districts and subprefectures, illustrated in Figure 3.3. The regionalization of the São Paulo map was also incorporated, as shown in Figure 3.4.

Considering that the official division of the city of São Paulo was created by law to guide the actions of municipal administration, each of these regions is highly independent, and they are divided in this way so that the entire city receives proper care in all aspects.

By leveraging these divisions, we gain a more granular understanding of mobility patterns within the intra-urban area under study. This level of detail is essential for

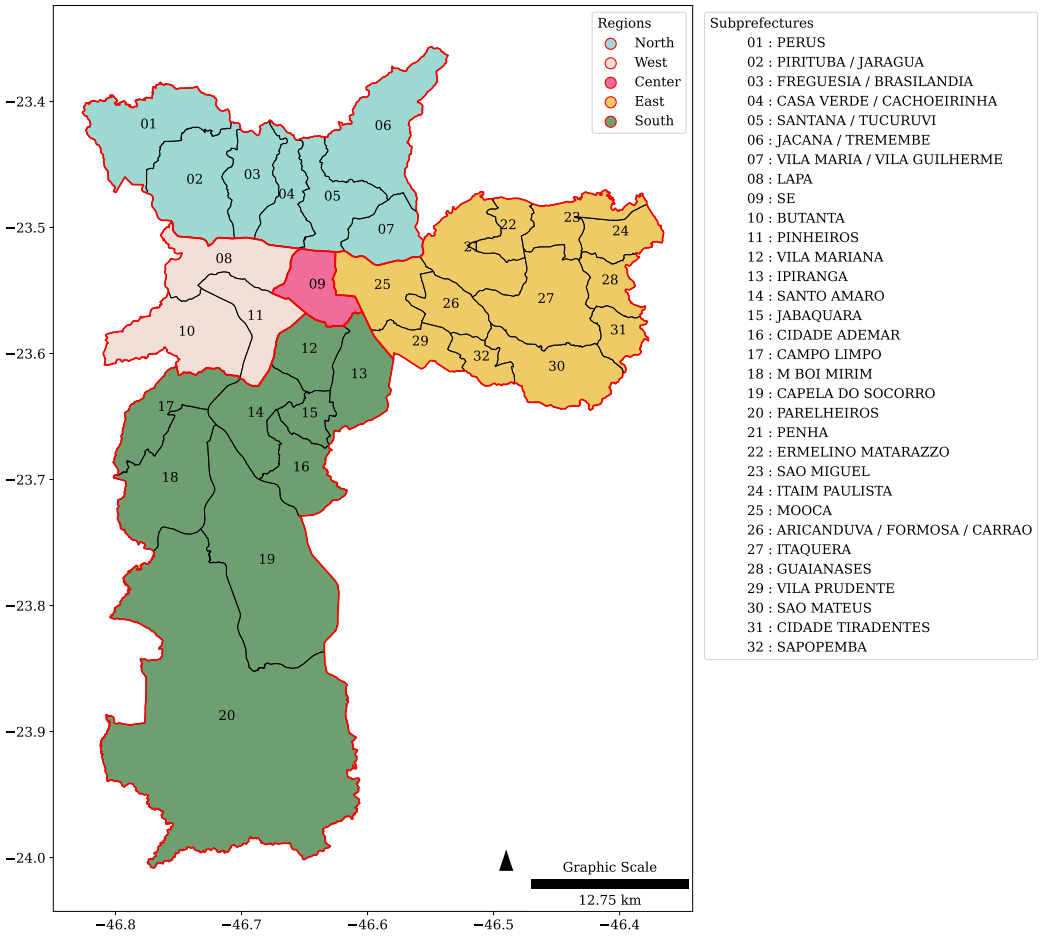
Figure 3.3 - Division of São Paulo's regions into sub-prefectures, districts, and zones.



accurately modeling the spread of infectious diseases, allowing us to identify high-traffic areas and potential hotspots for disease transmission. Additionally, examining these areas provides insights into socioeconomic factors impacting mobility patterns and, consequently, disease transmission rates.

Overall, integrating these divisions in our analysis enhances our comprehension of the intricate interplay between mobility patterns and disease transmission, offering critical insights for informing public health policy and response efforts.

Figure 3.4 - Regions of São Paulo.



3.2 Origin-Destination data

The OD Survey dataset utilized in our analysis spans 24 hours, commencing at 00:00 and concluding at 23:59 on the same day. Each individual in the dataset is assigned a unique identity number (ID PESS), and the dataset records the total number of trips made by each individual during that period (TOT VIAG). The dataset fields are summarized in Table 3.1.

Table 3.1 - Dataset attributes

Attribute	Meaning
ID PESS	Identity number of a person
N VIAG	Trip number performed by ID PESS
TOT VIAG	Total trips taken per person
ZONA O	Origin zone
ZONA D	Destination zone
H SAIDA	Departure hour from the ZONA O
MIN SAIDA	Departure minute from the ZONA O
H CHEG	Arrival hour from the ZONA D
MIN CHEG	Arrival minute from the ZONA D
CD-ATIVI	Type of activity carried out during the day
MODO1	Primary mode of transportation used

Individual trips are sequentially numbered (N VIAG) and include information on the origin zone (ZONE O) and destination zone (ZONE D) of the trip. The dataset also provides precise departure times from the origin zone (H SAIDA and MIN SAIDA) and arrival times at the destination zone (H CHEG and MIN CHEG). Table 3.2 illustrates the structure of the dataset, with the first row exemplifying a person (ID PESS) leaving at 5 hours (H SAIDA) and 45 minutes (MIN SAIDA) from an origin zone (ZONA O) coded as 1 and arriving at their destination zone (ZONA D) of number 3 at 5 hours (H CHEG) and 55 minutes (MIN CHEG). This trip is numbered 1 (N VIAG) out of a total of 2 trips (TOT VIAG) made that day.

While our study encompasses the entire city of São Paulo, it is noteworthy that in some zones among the designated 342, no households were visited during the survey. Due to the lower density or near absence of residences in these specific areas, they were excluded from the survey selection. To transparently address this, such instances will be documented as “Missing Values,” indicating which zones are not included in our study. This acknowledgment ensures clarity regarding the limitations

of our dataset, specifically identifying zones that may not contribute to the overall analysis due to a lack of residential representation.

Table 3.2 - Profile of the first 5 rows of the dataset of the city of Sao Paulo.

ID PESS	N VIAG	TOT VIAG	ZONA O	ZONA D	H SAIDA	MIN SAIDA	H CHEG	MIN CHEG
10001101	1	2	1	3	5	45	5	55
10001101	2	2	3	1	15	45	15	55
10001102	1	3	1	82	9	0	9	50
10001102	2	3	82	84	17	0	18	0
10001102	3	3	84	1	22	50	23	30

Significantly, journeys can occur within the same areas or extend across diverse regions.

3.3 Network construction

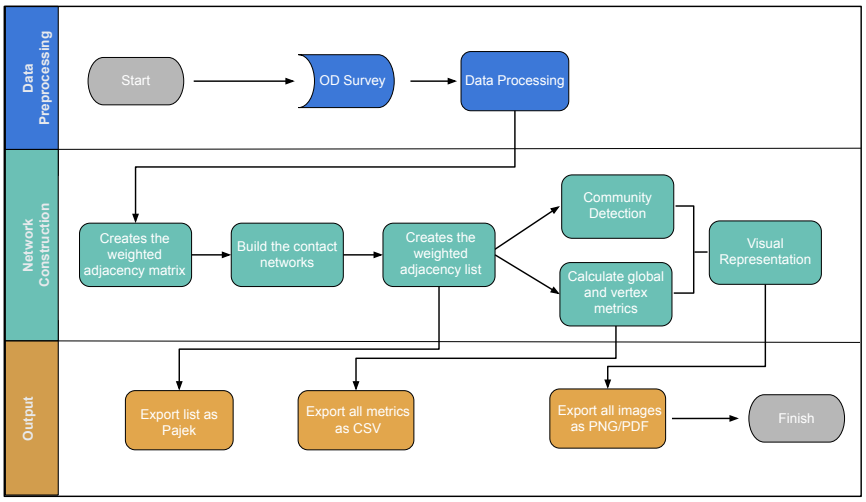
To process and analyze the Origin-Destination Survey dataset, we use the Python programming language along with the igraph (2023) library, a powerful tool for creating and manipulating graphs. The data underwent preprocessing and was subsequently exported as a CSV file.

To create all plots and graphics related to our network analysis, we chose to utilize the networkx library (HAGBERG et al., 2008). It proved more efficient in handling the extensive graph structure and enabled significantly faster image generation compared to alternative libraries.

The study's source code is organized into three main blocks: data preprocessing, contact network construction, and output. The flowchart in Figure 3.5 delineates these processes into distinct blocks.

In the first block, we tackled data inaccuracies within the OD survey by arranging them in chronological order based on each person's ID PESS and the number of trips. As part of the data processing, we also excluded individuals whose origin and destination zones were unspecified and those whose travel times were not registered. In the second block, individuals present in the same zones during the same periods were connected. We segmented the day into one-minute periods, identifying the zones each individual occupied at each minute, disregarding travel time between

Figure 3.5 - Flowcharts of all processes divided into 3 main blocks, representing respectively data preprocessing, contact network construction, and output.



zones. The resulting weighted adjacency matrix W represented the contact duration between individuals.

Within the second block, the weighted adjacency matrix was utilized to construct the contact network, $N = 39627$ and $L = 9418901$, and the weighted adjacency list. The adjacency list facilitated the calculation of metrics for each vertex, global network metrics, exploration of community structure, and generation of images.

In the third block, the results obtained from the contact network analysis were exported. The weighted adjacency list, created in the second step, was exported in PAJEK format, a standard format for storing and exchanging network data. This format comprises three columns, with the first and second columns representing the vertices and the third column representing the edge weight between them.

Following the export of the adjacency list, we computed and added the vertex and global metrics to their respective CSV files. By exporting the results in a standardized format and including both vertex and global metrics, we facilitate further analysis and comparison of the contact network with other networks or datasets.

An important point to mention is that all images generated for network visualization display only the vertices. In other words, each point corresponds to a vertex, and all these vertices have been placed within their respective residential zones, utilizing the São Paulo shapefile due to the network containing a vast number of edges, which would hinder the visualization of the entire graph.

4 RESULTS

The results chapter has been divided into three sections. The first section demonstrates the construction of the contact time list and the analysis of OD data without using the network. In the second section, the properties of the constructed network and its vertices are analyzed based on density, degree, strength, and clustering coefficient. Finally, in the third and last section, the analysis of communities are presented, including the distribution and map of the dominance coefficient. This section are also include the topological consistency in the map of the dominant community for each zone.

4.1 Data analysis and setting up the contact network

Firstly, before delving into the analysis of our network, it is necessary to demonstrate its construction. With an extensive dataset giving rise to our network of individuals, it is worthwhile to investigate some of the factors that shaped it.

The graph is built based on the contact time each person maintains with others throughout the day, and this contact is exclusively dependent on the shared zone in which these individuals are located. Each person (node) is placed in their residential zone, and this positioning will be maintained for all subsequent graph visualizations in this thesis.

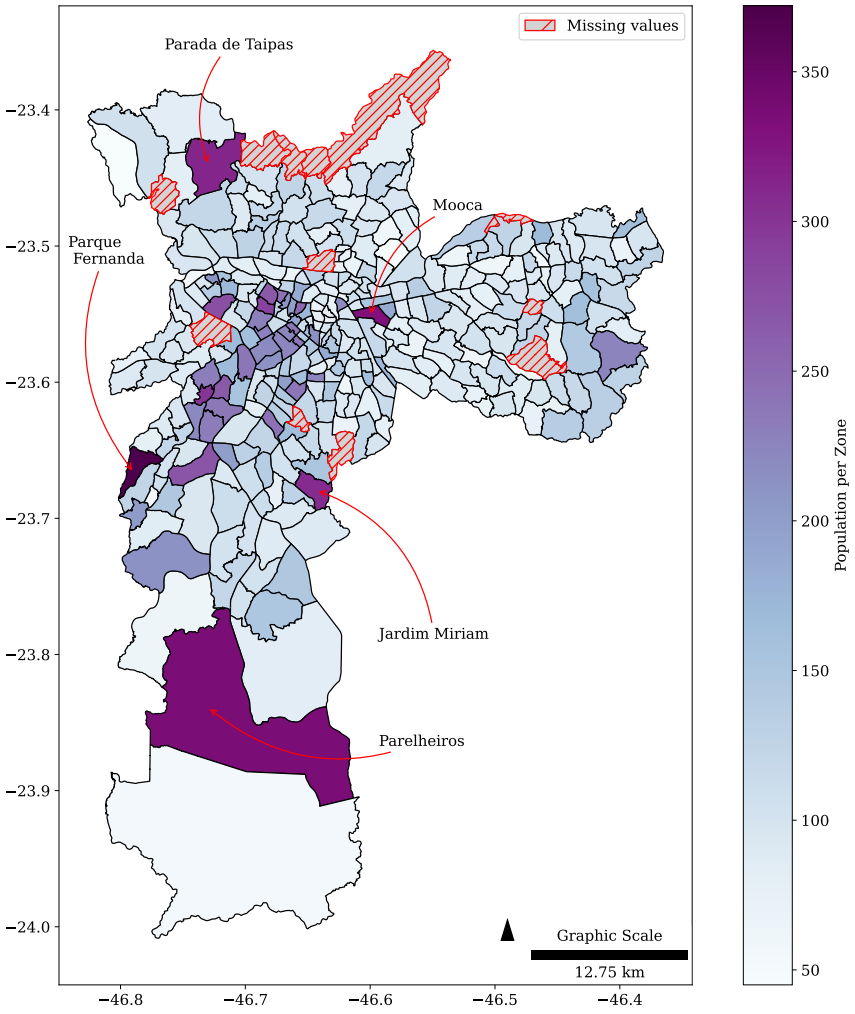
To illustrate the population distribution per zone, as depicted in Figure 4.1. By comparison, Figure 4.2 exhibits the population density of zones, which is a normalization of the population by area.

The central zones of the city of São Paulo, with an emphasis on São Carlos do Pinhal and Pamplona, maintain the highest population density despite not having as many people registered in the study as mentioned above.

By examining the departure time of each person, we obtain the number of trips per hour on a typical day within the city of São Paulo, as seen in Figure 4.3. However, these trips include both journeys made outside and inside the zone where each person was located.

This type of travel analysis was to be expected, considering that São Paulo is the busiest city in the entire metropolitan region (MARTINS et al., 2021). However, we can delve deeper into these trips and detect how many trips are made and how many zones are visited during this daily period (Figure 4.4).

Figure 4.1 - Number of people per residential zone.



According to Figure 4.4, the majority of people, approximately 70%, make two trips and visit two different zones in a single day. Combining this analysis with the histogram in Figure 4.3, we observe that most people undertake these trips during the periods between 06:00 and 07:00, around 12:00, and between 17:00 and 18:00.

With this data, we can reasonably infer that the majority of trips made by individuals on a typical day in the city of São Paulo are commuting trips. In other words, people depart from an origin zone (such as a residential zone), travel to a destination zone (a work or study zone, for example), and return to their origin zone in the early evening.

Figure 4.2 - Population density per residential zone.

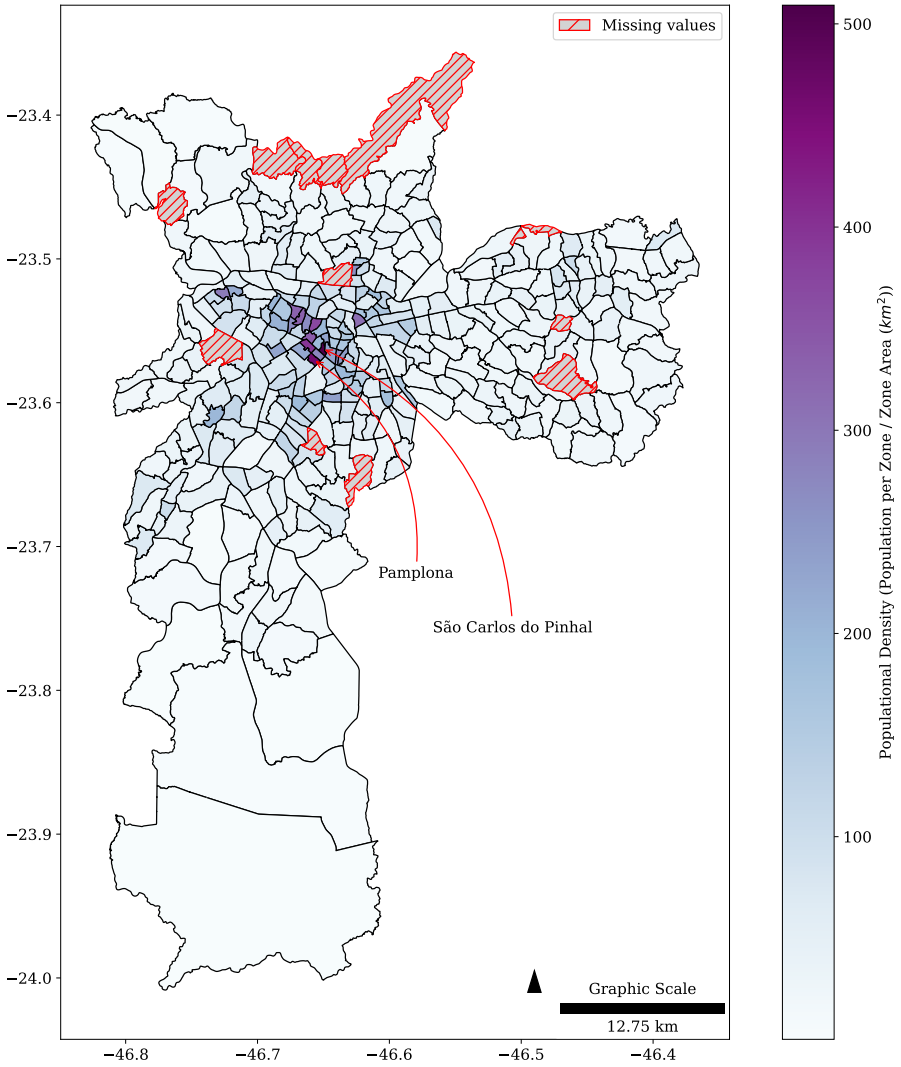


Figure 4.3 - Number of trips made at each hour of the day, starting from 00:00 and ending at 23:00.

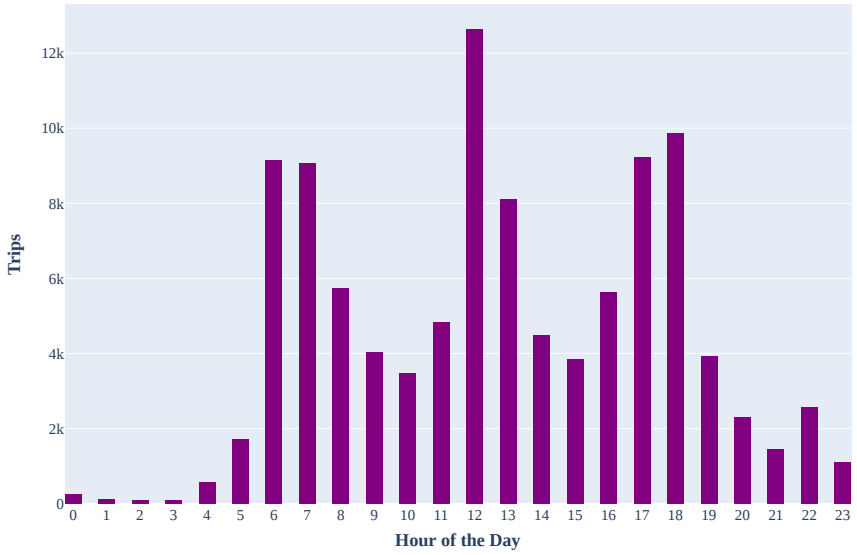
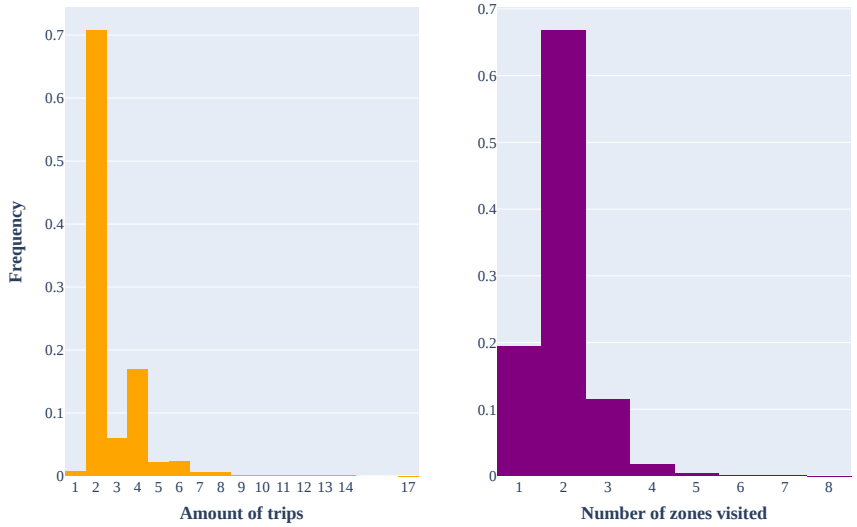


Figure 4.4 - Left: Frequency of the quantity of trips made; Right: Frequency of the number of visited zones. Both graphs represent a day's time span.



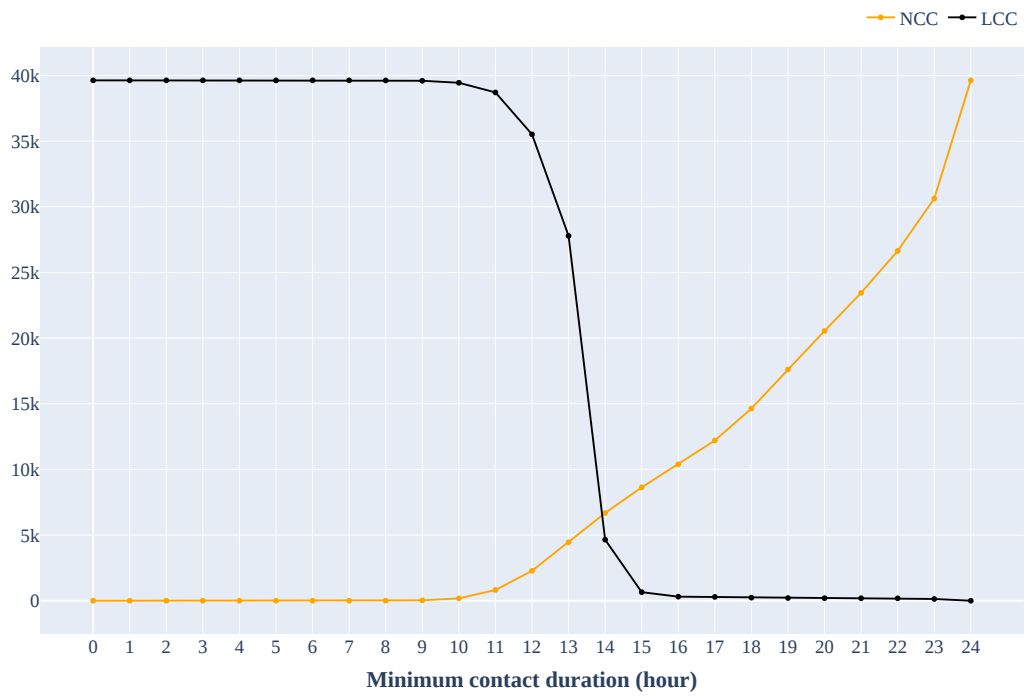
4.2 Network metrics

Between the 24 hours of Origin-Destination mobility data, we observe different minimum contact durations (mcd) to assess how they influence the structure of the contact network.

With the passing of each hour of the day, we have the threshold $mcd = 1, 2, 3, ...$, and edges with weights less than the specified number of hours will be cut from the network. The network in the initial state, without the passage of time, is considered with $mcd = 0$. For example, in the network with $mcd = 6$, all edges with weights less than 360 minutes are cut. In all subsequent results, the divisions of the mcd thresholds follow this procedure.

For each of the different values of mcd, we analyze the density, degree, strength, and clustering coefficient of the vertices that make up the network, as well as an average for the entire network. Additionally, we examine the connectivity of our network by analyzing the number of connected components and the most significant component in each case.

Figure 4.5 - The graph displays the number of connected components (NCC) and the largest connected component (LCC) for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$.



Initially, we analyze the connected components of the graph, as seen in Figure 4.5. As the *mcd* increases, it is expected that the links of the graph become disconnected, making the network increasingly disjoint over time. However, it is interesting to note that this network breakdown does not occur steadily. Notice that at $mcd = 12$, the network starts to lose connections much more rapidly, reaching $mcd = 24$ where all nodes in the network are entirely isolated, leaving us with the number of connected components (NCC) equal to the number of vertices, $N = 39627$.

This event can be observed in more detail in Figure 4.6. That illustrates the connectivity state of the network as the *mcd* increases. In the beginning, $mcd = 0$, all vertices are connected, and there is only one connected component. At $mcd = 8$, the first nodes start to detach from the most significant connected component. At $mcd = 16$, it shows the state of the network as the contact time progresses, and finally, at $mcd = 24$, the entire network is disconnected, and all nodes are isolated.

It is noteworthy that the nodes become disconnected uniformly throughout the city of São Paulo, regardless of the zone or region. This observation can be related to the information in Figure 4.3. According to our study, the network ceases to have only one connected component at the time of $mcd = 7$, where in Figure 4.3, it was the first peak in the number of trips during the day. Also, when the network starts to disconnect more rapidly at the time of $mcd = 12$, it corresponds to the highest number of trips. Indicates that the loss of connections is indeed associated with the trips people make during the day.

In Figure 4.7, we have the average degree $\langle k \rangle$ of the graph represented for each point, and as the *mcd* increases, the average degree decreases. Its behavior is identical to that of density, as seen in Figure 4.8. An expected behavior is because as edges are excluded from the network, the number of connections for each vertex also decreases.

The average strength of the network $\langle s \rangle$, however, exhibits a different behavior, as seen in Figure 4.9. As the *mcd* increases and reaches the 12-hour mark, a change in the curve's shape is noticeable. A non-linear regression asserts that it exhibits a sigmoidal behavior, given by the equation $y = \frac{3622.90}{(1+0.23e^{(x-10.77)})} - 62.37$. In other words, from the 12-hour mark onward, the strength of the graph starts to decline more rapidly.

Similar to the average strength of the graph $\langle s \rangle$, the clustering coefficient $\langle c \rangle$ is also affected by the large number of trips made at 12 hours, as seen in Figure 4.10. This behavior can be explained because, as the *mcd* increases, weaker links are

Figure 4.6 - This figure shows the increase in NCC as the mcd increases. Blue nodes are still connected to the network while red are isolated.

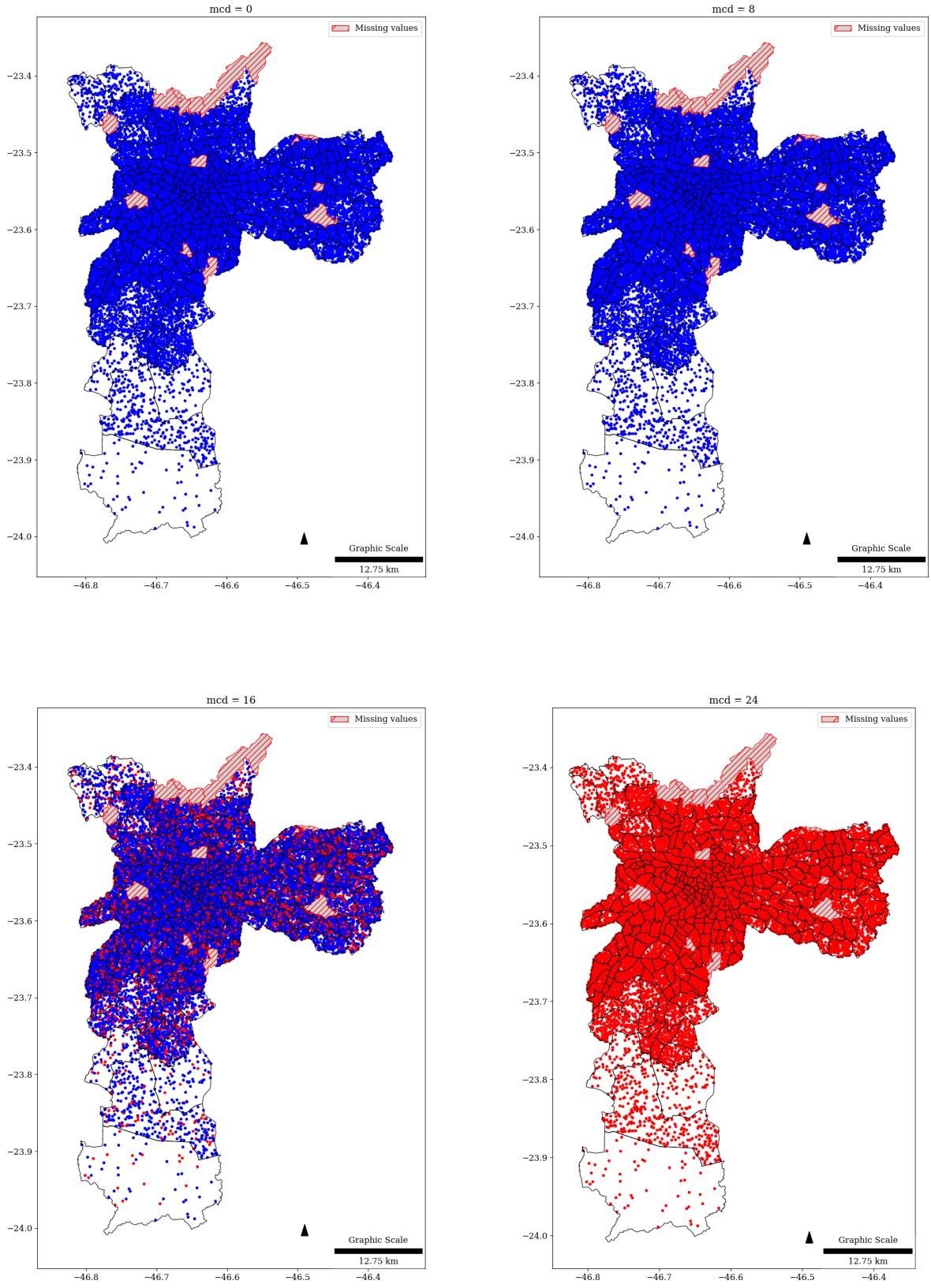
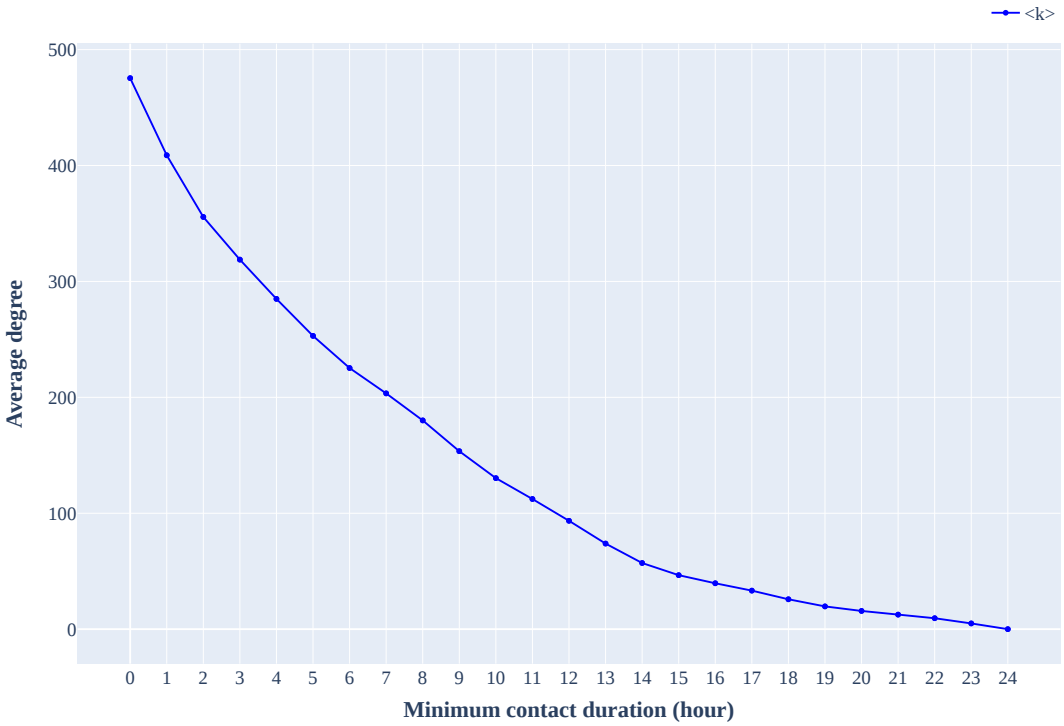


Figure 4.7 - The graph displays the average degree for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the average degree $\langle k \rangle$ of a graph generated with the corresponding threshold value.

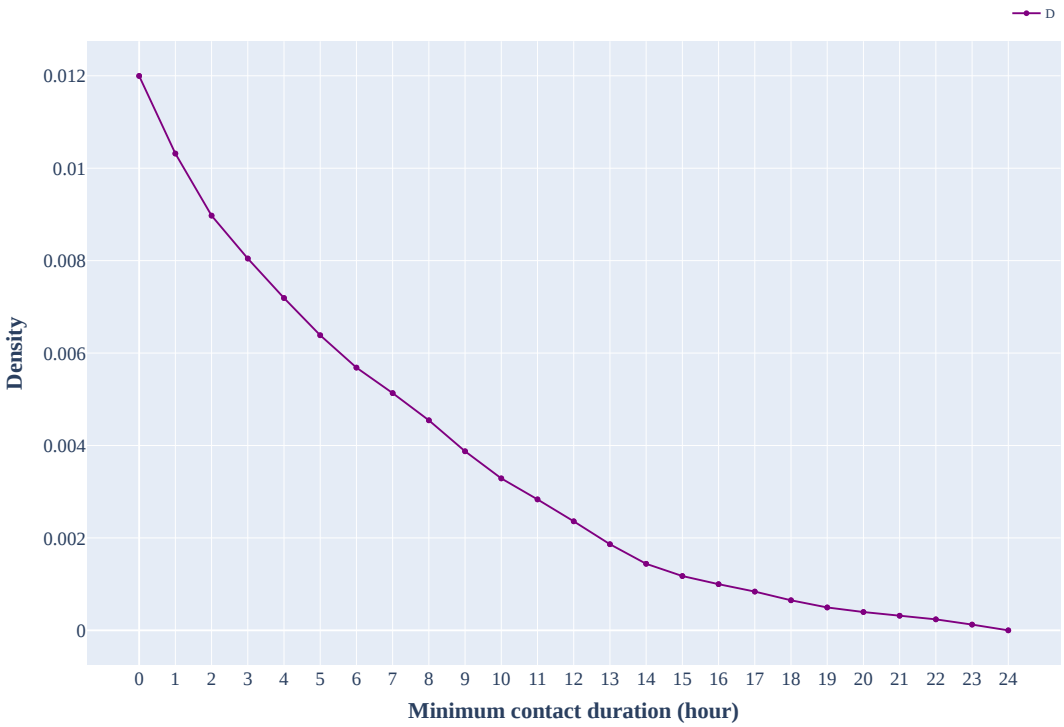


disconnected, resulting in a more clustered graph, with neighbors of a node also being neighbors themselves. However, from 12 hours onward, the coefficient begins to decline drastically, indicating that a significant portion of these clusters has a contact strength close to 12 hours, as also shown in Figure 4.9.

Furthermore, in Figure 4.6, we observe that nodes are disconnected uniformly across the entire study area. It suggests that there is no larger cluster in any specific region of São Paulo, and clusters are small groups scattered throughout the city.

In addition to the (global) metrics calculated for the entire graph, we also compute local metrics for the nodes. Figures 4.11, 4.12, and 4.13 depict histograms for each of the presented *mcd* values in the study. For each histogram, the x-axis is fixed, and the values on the y-axis are normalized to make changes in each variable more readable.

Figure 4.8 - The graph displays the density for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the density D of a graph generated with the corresponding threshold value.



Similar to the trends observed in $\langle k \rangle$, $\langle s \rangle$, and $\langle c \rangle$, here, one can discern with more detail the influence of the high number of trips at 12 hours. From $mcd = 12$ onward, the metric values for each vertex start decreasing more rapidly, as indicated by the swift leftward shift in the histograms.

Finally, an analysis of the IDs of individuals is conducted based on the calculated metrics, as shown in Table 4.1. Initially, the top 100 nodes with the highest and lowest degree, strength, and clustering in the default graph ($mcd = 0$) are extracted. Subsequently, two additional measures are analyzed, as provided by the 2017 OD Survey, as seen in Table 3.1.

For these measures, the most common categories were selected according to the segments of the individuals surveyed.

Figure 4.9 - The graph displays the average strength for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the average strength $\langle s \rangle$ of a graph generated with the corresponding threshold value.

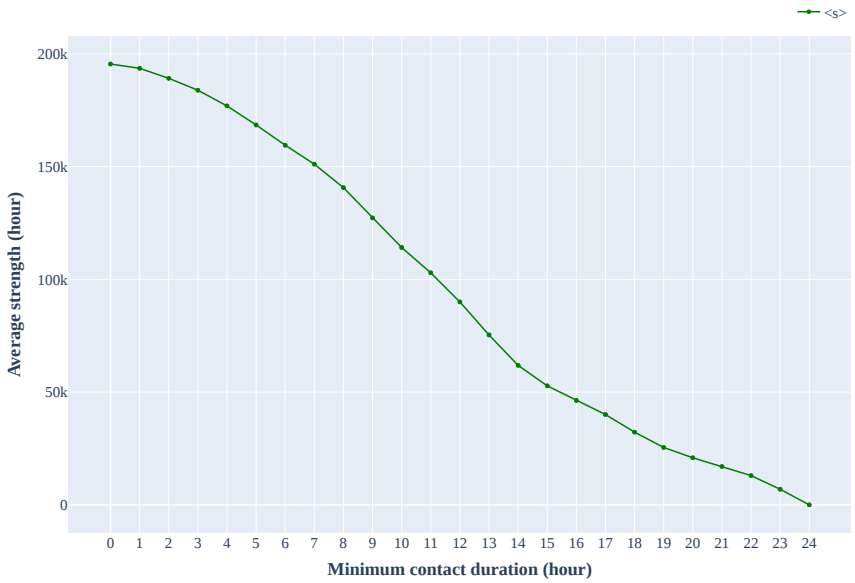


Table 4.1 - The most common activity performed and primary mode of transportation.

	CD-ATIVI	MODO1
Highest Degree	Has regular job (81%)	Driving a Car (41%)
Lowest Degree	Student (45.5%)	On Foot (66.3%)
Highest Strength	Has regular job (84%)	On Foot (36%)
Lowest Strength	Has regular job (41%)	On Foot (40%)
Highest Clustering	Student (33.7%)	On Foot (79.2%)
Lowest Clustering	Has regular job (57.4%)	Driving Car (40.6%)

Figure 4.10 - The graph displays the average strength for a series of threshold values ranging from $mcd = 0$ to $mcd = 24$. Each point on the graph represents the average clustering $\langle c \rangle$ of a graph generated with the corresponding threshold value.

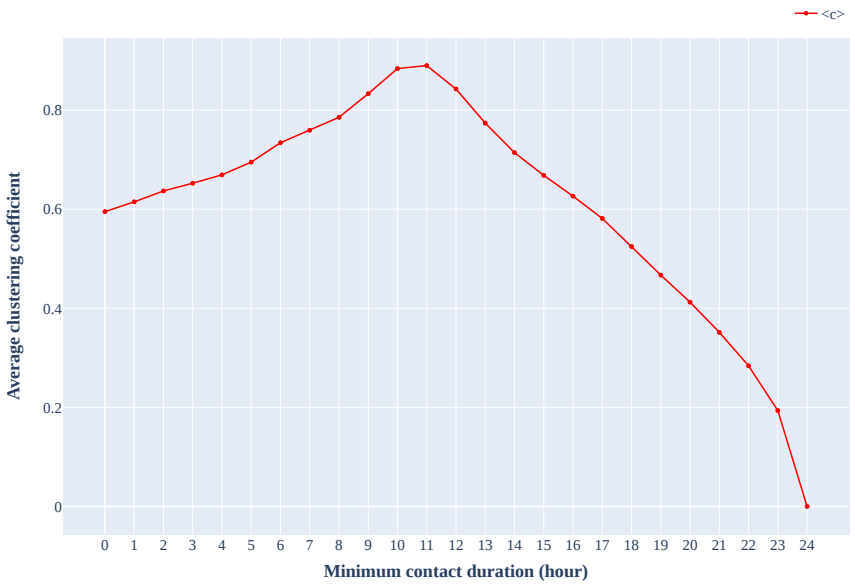


Figure 4.11 - In this figure, we have 24 graphs, each one representing the degree of the vertices of the network. Considering that each point represents a vertex, the y-axis contains the number of vertices with the given degree, represented by the x-axis.

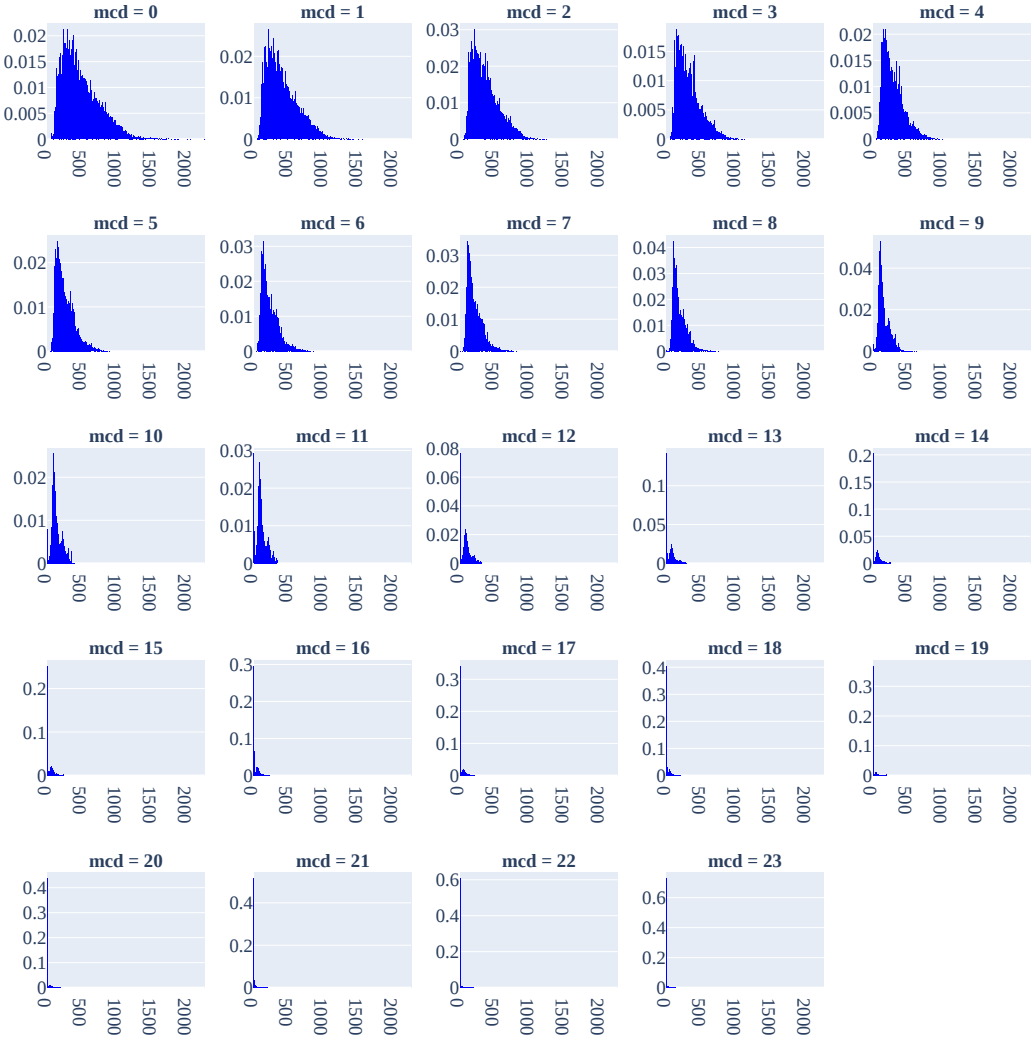


Figure 4.12 - In this figure, we have 24 graphs, each one representing the strenght of the vertices of the network. Considering that each point represents a vertex, the y-axis contains the number of vertices with the given strenght, represented by the x-axis.

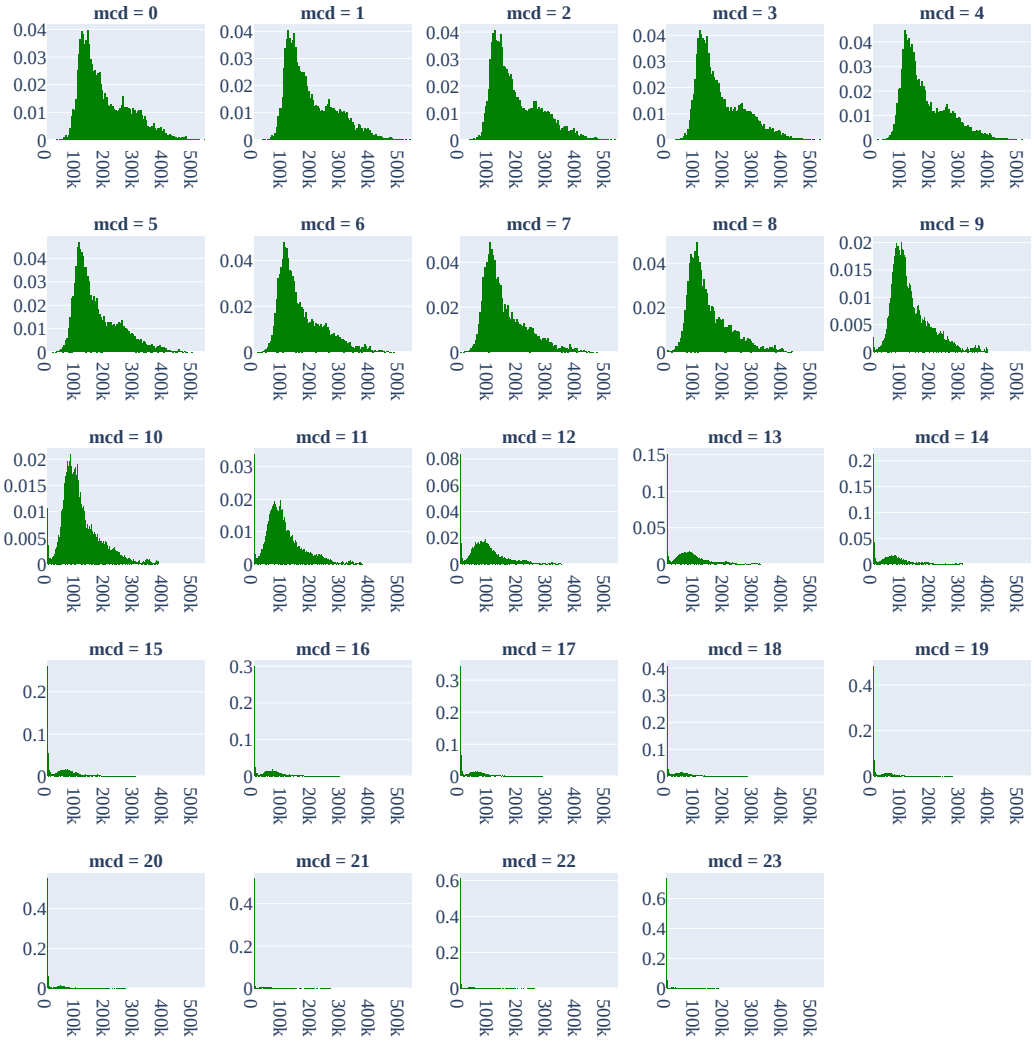
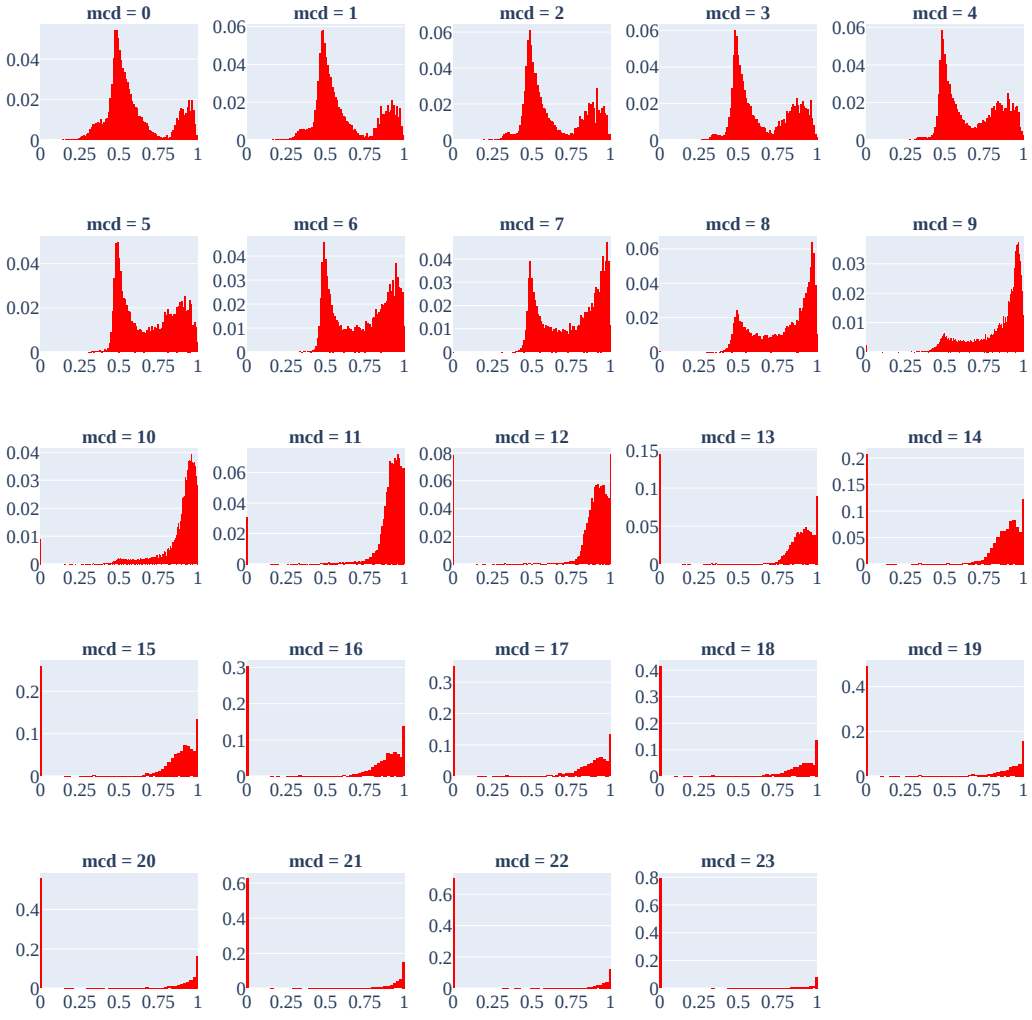


Figure 4.13 - In this figure, we have 24 graphs, each one representing the clustering of the vertices of the network. Considering that each point represents a vertex, the y-axis contains the number of vertices with the given clustering, represented by the x-axis.



4.2.1 Ordinary and outer strength

After analyzing all these metrics, the network is based on the shapefile of São Paulo to analyze the regions with higher strength. For this, the strength of each vertex is summed for each residential zone, varying the *mcd* (Figure 4.14).

It is noticeable that the most populous zones, as shown in Figure 4.1, have the highest strengths, as expected. However, for comparison, we also compute the strength for each vertex, including only connections to vertices of different residential zones, which we call the outer strength (Figure 4.15). Similar to Equation (2.4), it is possible to define an equation for the outer strength formally:

$$s_i^{(l)} = \sum_{j=1}^{N} a_{ij}^{(m)} w_{ij}^{(m)}, (m \neq l); \tag{4.1}$$

where the outer strength ($s_i^{(l)}$) of a node i in a zone l will be the sum of the weights of the edges of its connections, provided that node j is in a zone m, where $m \neq l$.

Comparing Figures 4.14 and 4.15, one observes that the zones farther from the center lost strength when vertices from the same residential zone are disregarded. It indicates that a significant portion of nodes in the more distant zones of the city remain within their residential zone for much of the day.

Figure 4.14 - The strength of each vertex varies with the *mcd*.

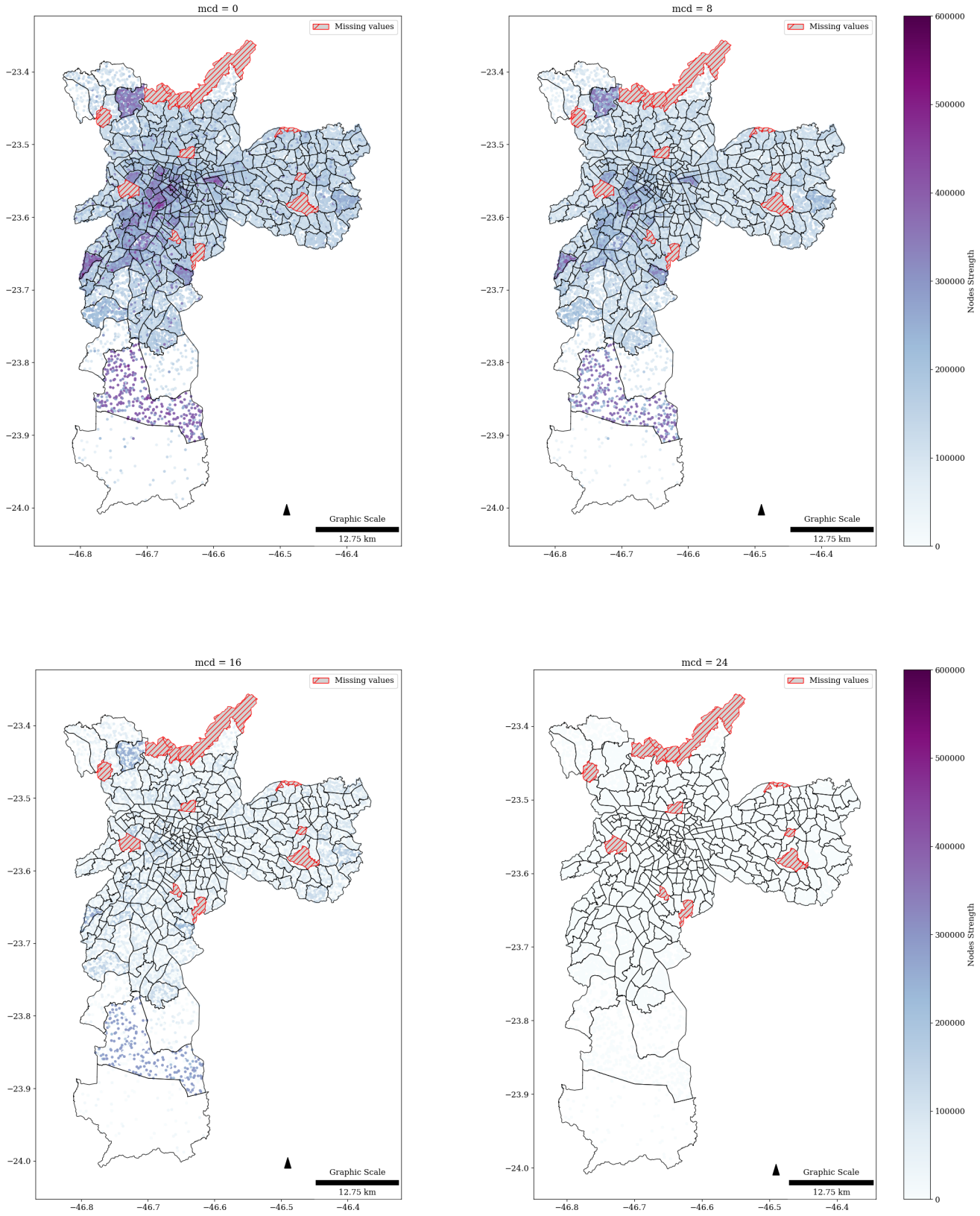
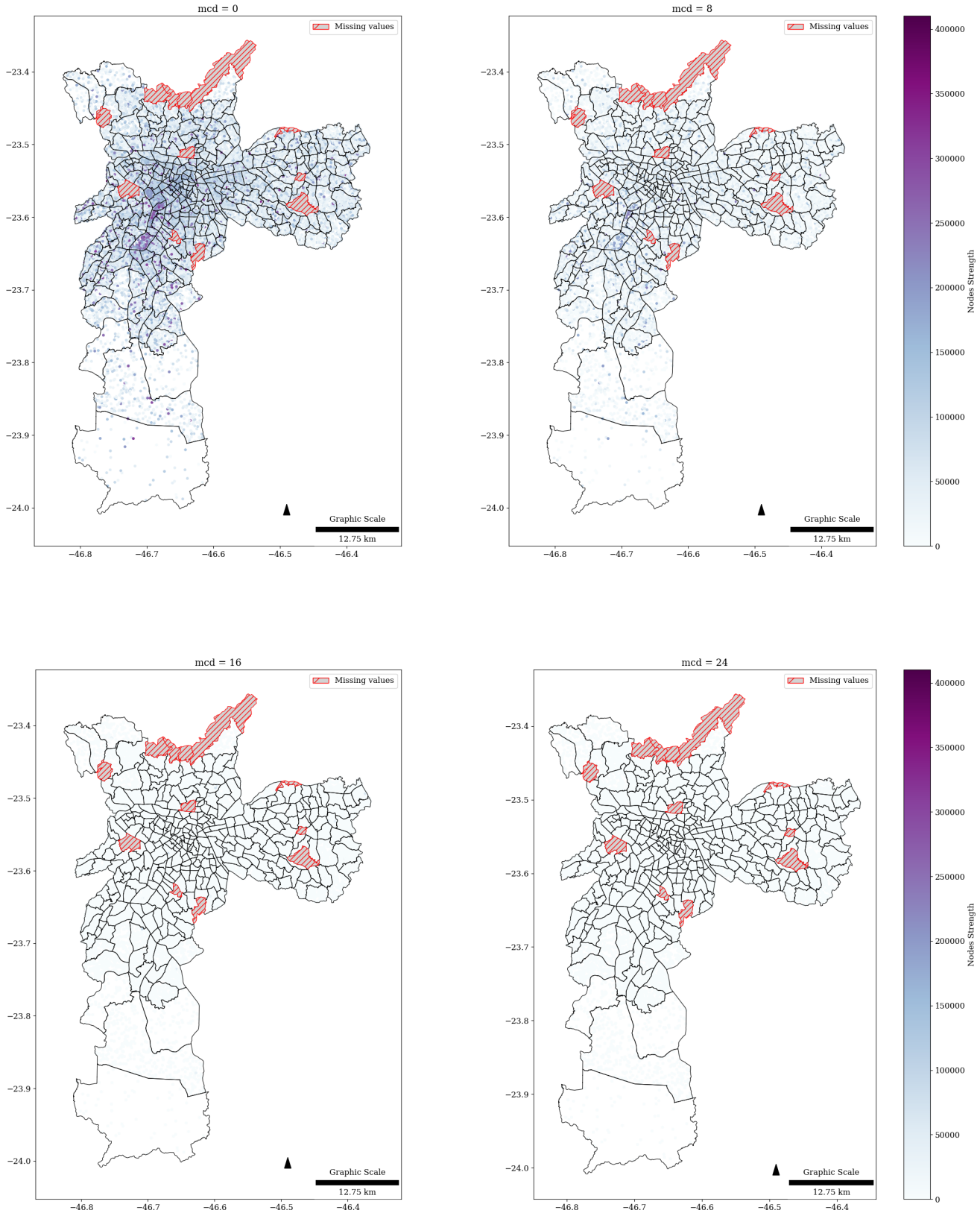


Figure 4.15 - The outer strength of each vertex varies with the *mcd*.



4.3 Community detection

The third section of results starts with the study of communities. As mentioned in Chapter 2, we use the Louvain algorithm for community detection. As the Louvain method operates based on a resolution, the main challenge is to find a resolution that explains the case under study.

We conducted several tests to find the best resolution that explains a meaningful way to divide the contact network for the city of São Paulo for $mcd = 0$. For instance, with the default parameter of $r = 1$, the code returns 26 communities, which do not provide a clear physical explanation of the division despite its modular value still being relatively high $Q = 0.67$.

The main goal here is to find a community division that makes sense with the explored data until then. For example, with $r = 0.3$, we find four different communities as shown in Figure 4.16.

Similar to the other presented figures, each point represents a node in the network, and its color indicates which community the respective vertex belongs to. Noticeably, zones do not have vertices belonging to only one community, so we determine the percentage of the most frequent for each zone to assess how much a community dominates a specific area. Figures 4.18 and 4.17 demonstrate that all zones are well divided among the respective communities, with a significant portion having a community representing more than 90% of its population.

We divide each zone by the dominant community that represents it, thus providing a better visual representation of the communities for the city of São Paulo as a whole. This representation can be visualized in Figure 4.19.

The corresponding modularity for the chosen resolution of $r = 0.3$ is $Q = 0.76$, which is even better than the one obtained with the default value of $r = 1$. This division closely resembles the division of city regions presented in Figure 3.4.

The comparison between the identified communities with the São Paulo region is presented in Table 4.2. Dividing the regions of São Paulo, Figure 3.4, into zones, we have the following distribution: the North region has 50 zones, the East region has 100 zones, the combination of the West and Central regions has 90 zones, and the South region has 89 zones.

Figure 4.16 - Each point represents a node, and its color corresponds to each of the communities presented.

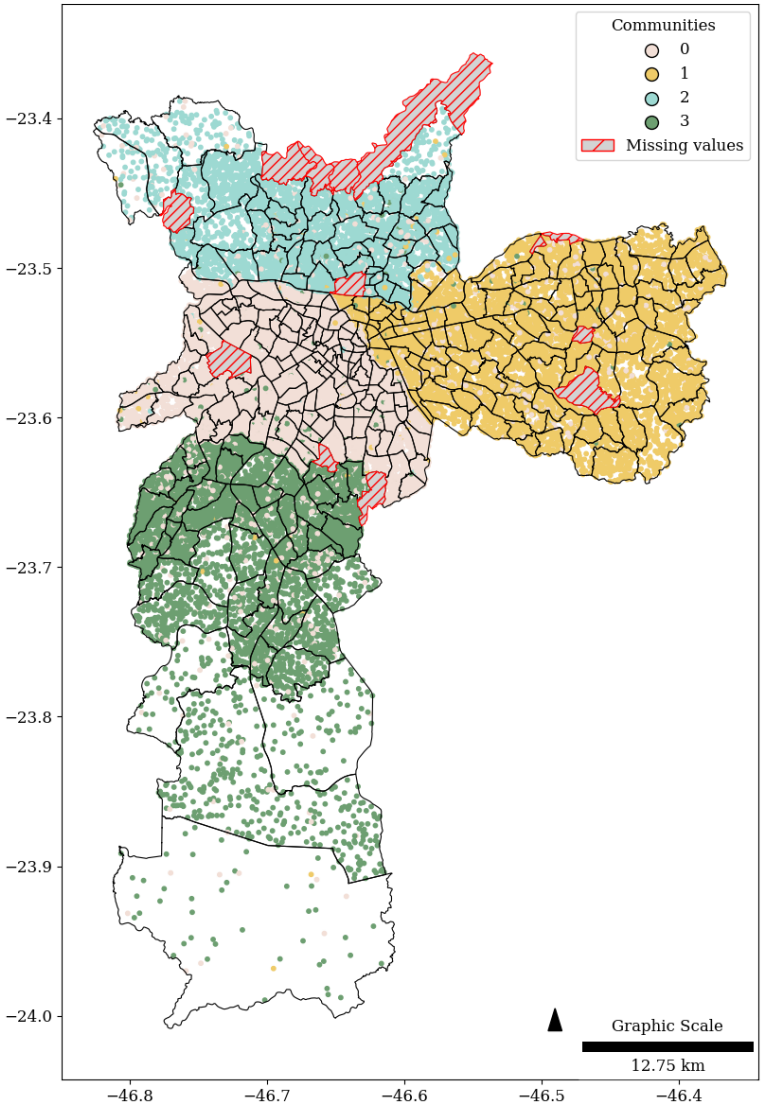


Figure 4.17 - Dominance of each zone.

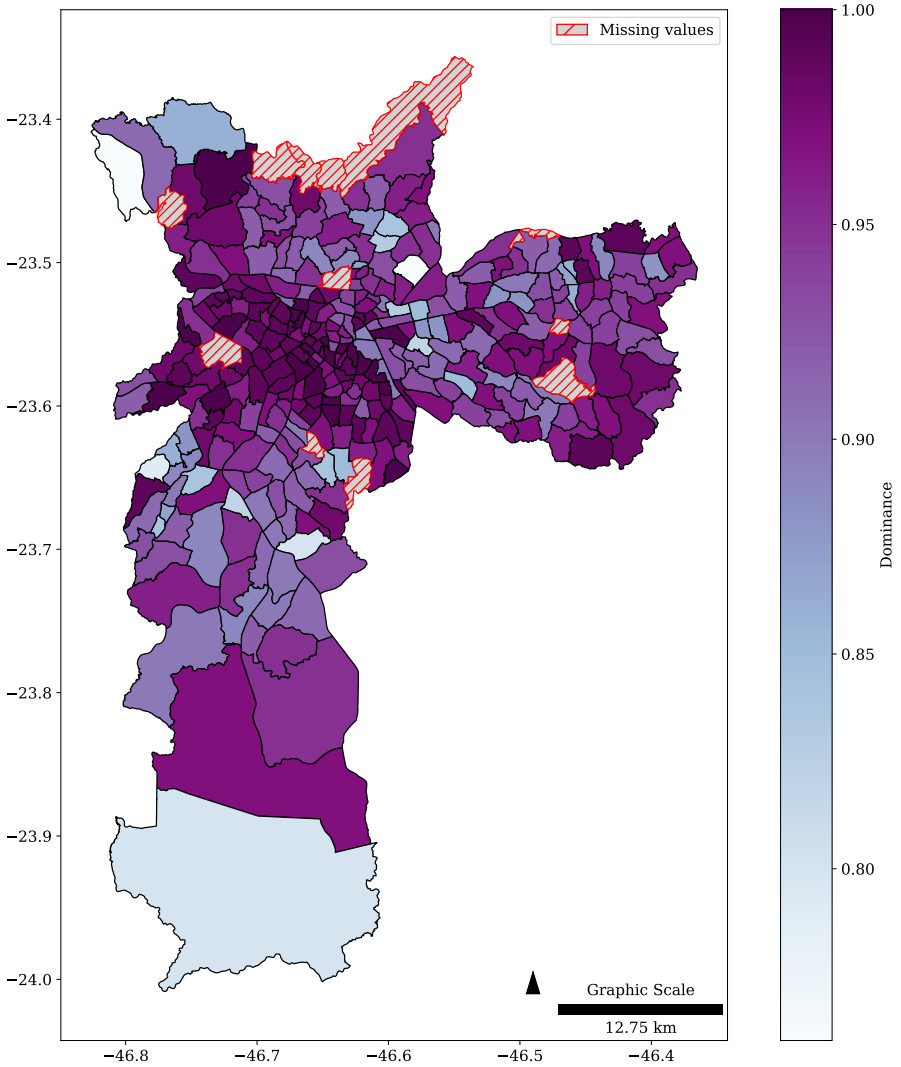
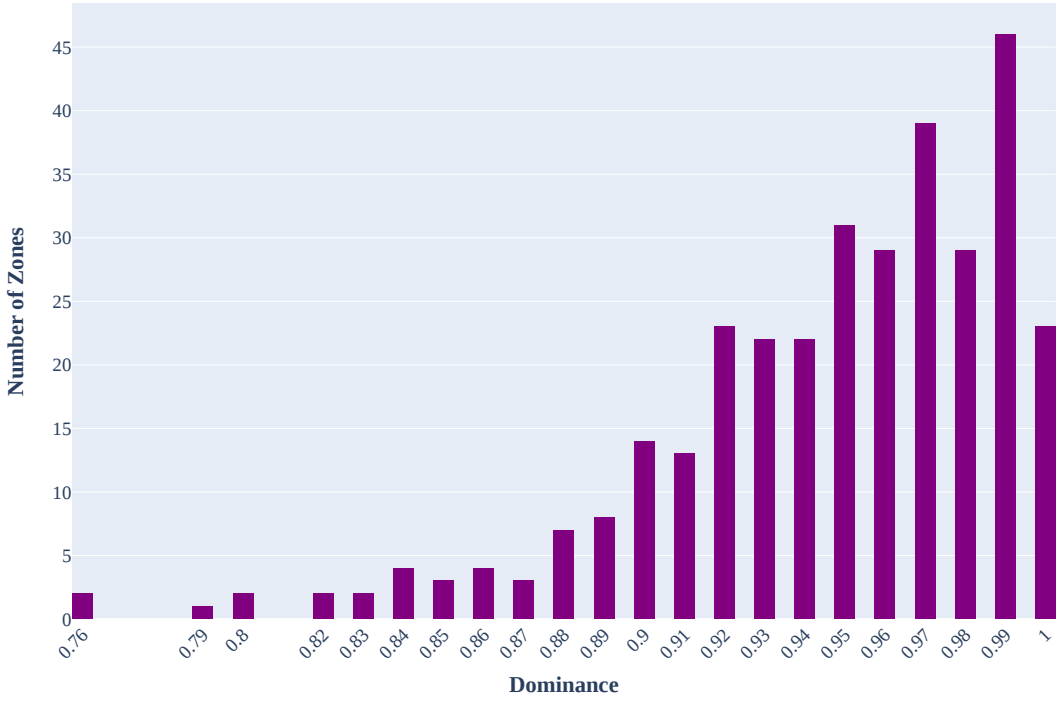


Figure 4.18 - Histogram of dominance values.



Examining Community 2, which represents the North region with 47 zones out of the total 50, reveals a deficit of 3 zones. This places Community 2 at a shortfall in comparison to Community 0, which is associated with the Central/West region.

Community 1, representing the East region, demonstrates a surplus of 4 zones, 3 zones in comparison to Community 0 and 1 zone relative to Community 3. Shifting the focus to Community 0, representing the Central/West region, it experiences a deficiency of 3 zones in comparison to Community 1 while maintaining an excess of 3 zones compared to Community 2. It also has a surplus of 38 zones in relation to Community 3.

Lastly, Community 3, representing the South region, faces a deficit of 38 zones compared to Community 0 and falls short by 1 zone compared to Community 1. This analysis provides information on distribution disparities between the identified communities and their respective regions. It also emphasizes the more significant division observed in the North and East regions while the Center and West regions converge, subtly merging even with the South region.

Figure 4.19 - Community division by zones.

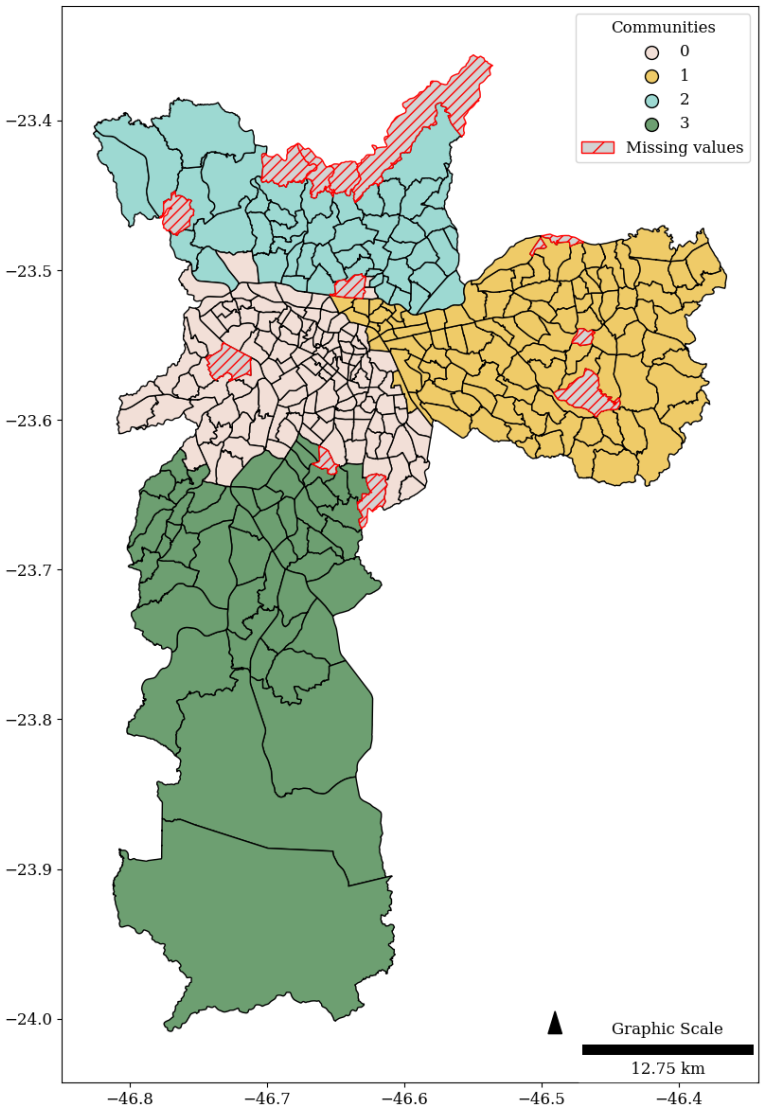


Table 4.2 - Comparison of Zone Distributions among Regions and Communities in São Paulo.

Region/ Community	Number of zones per region	Number of zones per community	Surplus(+)/ Deficit(−)
North	50	47	−3 zones from Community 0
East	100	104	+3 zones from Community 0 +1 zones from Community 3
Center and West	90	128	−3 zones from Community 1 +3 zones from Community 2 +38 zones from Community 3
South	89	50	−1 zones from Community 1 −38 zones from Community 0

5 CONCLUSIONS

In this thesis, we study the structure of the contact network derived from mobility data for the city of São Paulo. By unifying this structure with the residential components of the vertices, we can track visually how the contact network changes over time.

As the minimum contact duration (*mcd*) changes, there is an expected decrease in centrality and clustering coefficient values. Initially, the network is highly connected, and as *mcd* increases, edges that do not reach this value are removed.

However, with the progression of *mcd*, reaching a peak at 07:00, the network begins to transition from being complete to having isolated nodes. At the 12:00 mark, it undergoes a complete behavior change, indicating critical times of movement during the day.

Regarding people's relationships between the places they travel during the day, despite having regular activities and moving to other zones, the factor that most impacts these links is the residential zones.

Another factor demonstrating this is the positions of the connected components. With the increase in *mcd* and the growth of connected components, isolated vertices are seen to increase uniformly throughout São Paulo, negating the idea that people in the center would be more strongly connected.

It is plausible to observe the configuration taken by the four different generated communities. Considering the independence of each city region, the community structure respects this regional division to some extent.

Communities 1 and 2 represent the East and North regions well, respectively. Community 0 is mixed between the Central and West regions, and Community 3 represents part of the South region. It shows that the community structure follows residential components.

The proposed objectives for this dissertation are being achieved, but some limitations are encountered during the project. The limitations inherent in this study primarily stem from the constraints related to the available data. A more extensive temporal scope beyond the confined 24-hour window would have enhanced the depth of our research.

While our study is centered on understanding epidemic transmission dynamics and analyzing risk scenarios, it is crucial to acknowledge that our data and study zone are at a broader scale. As a prospective avenue for future research, there is merit in transitioning to a more localized perspective, honing in on a specific community. This approach could offer valuable insights into how a localized community's dynamics influence the broader structural framework of the city. Moreover, simulations involving agents, particularly intra-zone movement, could significantly enhance our understanding of these dynamics.

Regarding communities, the Louvain method for detection is a versatile algorithm that focuses on modularity optimization. Due to working with an extensive graph, the choice for the method is almost exclusively due to its meager computational cost compared to other models used in the literature. It suggests the use of adaptations of this method (BLONDEL et al., 2023) for future comparisons.

Finally, in addition to these improvements and the study of disease transmission such as COVID-19, the goal is to analyze disaster risk scenarios based on the analysis of this complex network.

REFERENCES

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

ALLEN, L. J.; BRAUER, F.; DRIESSCHE, P. Van den; WU, J. **Mathematical epidemiology**. [S.l.]: Springer, 2008. ISSN 0075-8434. 1

BANSAL, S.; READ, J.; POURBOHLOUL, B.; MEYERS, L. A. The dynamic nature of contact networks in infectious disease epidemiology. **Journal of Biological Dynamics**, v. 4, n. 5, p. 478–489, 2010. 1

BARABáSI, A.-L.; PóSFAI, M. **Network science**. Cambridge: Cambridge University Press, 2016. ISBN 9781107076266 1107076269. Available from: <<http://barabasi.com/networksciencebook/>>. 6, 7, 8, 9

BARRAT, A.; BARTHELEMY, M.; PASTOR-SATORRAS, R.; VESPIGNANI, A. The architecture of complex weighted networks. **Proceedings of the National Academy of Sciences**, v. 101, n. 11, p. 3747–3752, 2004. 8

BLONDEL, V.; GUILLAUME, J.-L.; LAMBIOTTE, R. Fast unfolding of communities in large networks: 15 years later. **arXiv preprint arXiv:2311.06047**, 2023. 11, 48

BLONDEL, V. D.; GUILLAUME, J.-L.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. **Journal of Statistical Mechanics: Theory and Experiment**, v. 2008, n. 10, p. P10008, 2008. 9, 11

BOAVENTURA NETTO, P. O.; JURKIEWICZ, S. **Grafos: introdução e prática**. [S.l.]: Blucher, 2017. ISBN 9788521211334. 5

BONA, A. D.; FONSECA, K.; ROSA, M.; LÜDERS, R.; DELGADO, M. Analysis of public bus transportation of a brazilian city based on the theory of complex networks using the p-space. **Mathematical Problems in Engineering**, v. 2016, 2016. 1

CORREIA, M. d. M. G. et al. Grafo bipartido para análise das relações entre pessoas e lugares. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 9, n. 1, 2022. 3

_____. Study of links between people in urban areas based on mobility data for the city of São Paulo. **Proceeding Series of the Brazilian Society of Computational and Applied Mathematics**, v. 10, n. 1, p. 2–7, 2023. 3

COSTA, L. d. F.; JR, O. N. O.; TRAVIESO, G.; RODRIGUES, F. A.; BOAS, P. R. V.; ANTIQUEIRA, L.; VIANA, M. P.; ROCHA, L. E. C. Analyzing and modeling real-world phenomena with complex networks: a survey of applications. **Advances in Physics**, v. 60, n. 3, p. 329–412, 2011. 1

COTA, W. et al. Monitoring the number of Covid-19 cases and deaths in Brazil at municipal and federative units level. **SciELO Preprints**, 2020. 1

FORTUNATO, S. Community detection in graphs. **Physics Reports**, v. 486, n. 3-5, p. 75–174, 2010. 9

FREITAS, V. L.; MOREIRA, G. J.; SANTOS, L. B. Robustness analysis in an inter-cities mobility network: modeling municipal, state and federal initiatives as failures and attacks toward Sars-Cov-2 containment. **PeerJ**, v. 8, p. e10287, 2020. ISSN 2167-8359. 1

FREITAS, V. L. d. S.; KONSTANTYNER, T. C. R. d. O.; MENDES, J. F.; SEPETAUSKAS, C. S. d. N.; SANTOS, L. B. L. The correspondence between the structure of the terrestrial mobility network and the spreading of Covid-19 in Brazil. **Cadernos de Saúde Pública**, v. 36, p. e00184820, 2020. ISSN 1678-4464. 1, 12

GEOSAMPA. **Catálogo de metadados**. 2017. Available from: <https://metadados.geosampa.prefeitura.sp.gov.br/geonetwork/srv/por/catalog.search#/home>. Access on: 22 Jan. 2024. 16

GOEL, R.; BONNETAIN, L.; SHARMA, R.; FURNO, A. Mobility-based sir model for complex networks: with case study of Covid-19. **Social Network Analysis and Mining**, v. 11, p. 1–18, 2021. 12

GONZALEZ, M. C.; HIDALGO, C. A.; BARABASI, A.-L. Understanding individual human mobility patterns. **Nature**, v. 453, n. 7196, p. 779–782, 2008. 1

HAGBERG, A. A.; SCHULT, D. A.; SWART, P. J. Exploring network structure, dynamics, and function using networkx. In: VAROQUAUX, G.; VAUGHT, T.; MILLMAN, J. (Ed.). **Proceedings of the 7th Python in Science Conference**. Pasadena, CA USA: [s.n.], 2008. p. 11 – 15. 20

HÂNCEAN, M.-G.; SLAVINEC, M.; PERC, M. The impact of human mobility networks on the global spread of Covid-19. **Journal of Complex Networks**, v. 8, n. 6, p. cnaa041, 2020. 12

HARTNETT, G. S.; PARKER, E.; GULDEN, T. R.; VARDAVAS, R.; KRAVITZ, D. Modelling the impact of social distancing and targeted vaccination on the spread of Covid-19 through a real city-scale contact network. **Journal of Complex Networks**, v. 9, n. 6, p. cnab042, 2021. 12

IGRAPH. **igraph**. Zenodo, 2023. Available from: <<https://igraph.org/>>. 20

INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS. **Plano Diretor do INPE 2022-2026**, São José dos Campos: INPE, 2022. 3

LAMOSA, J. D.; TOMAS, L. R.; QUILES, M. G.; LONDE, L. R.; SANTOS, L. B.; MACAU, E. E. Topological indexes and community structure for urban mobility networks: variations in a business day. **PLoS ONE**, v. 16, n. 3, p. e0248126, 2021. ISSN 19326203. 1, 12

LIU, S.-Y.; BARONCHELLI, A.; PERRA, N. Contagion dynamics in time-varying metapopulation networks. **Physical Review E**, v. 87, n. 3, p. 032805, 2013. 12

MARTINS, T. G.; LAGO, N.; SANTANA, E. F.; TELEA, A.; KON, F.; SOUZA, H. A. de. Using bundling to visualize multivariate urban mobility structure patterns in the São Paulo metropolitan area. **Journal of Internet Services and Applications**, v. 12, p. 1–32, 2021. 13, 23

NEWMAN, M. E. Detecting community structure in networks. **The European Physical Journal B**, v. 38, p. 321–330, 2004a. 9

_____. Analysis of weighted networks. **Physical Review E**, v. 70, n. 5, p. 056131, 2004b. 10

NEWMAN, M. E.; GIRVAN, M. Finding and evaluating community structure in networks. **Physical Review E**, v. 69, n. 2, p. 026113, 2004. 9

OD SURVEY. **Pesquisa origem e destino**. 2017. Available from: <https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino>. Access on: 22 Jan. 2024. 2, 15, 16

PASTOR-SATORRAS, R.; CASTELLANO, C.; MIEGHEM, P. V.; VESPIGNANI, A. Epidemic processes in complex networks. **Reviews of Modern Physics**, v. 87, n. 3, p. 925, 2015. 12

PECHLIVANOGLOU, T.; LI, J.; SUN, J.; HEIDARI, F.; PAPAGELIS, M. Epidemic spreading in trajectory networks. **Big Data Research**, v. 27, p. 100275, 2022. 12

PINTO, E. R.; NEPOMUCENO, E. G.; CAMPANHARO, A. S. Impact of network topology on the spread of infectious diseases. **TEMA (São Carlos)**, v. 21, p. 95–115, 2020. 1

SANTOS, L. B.; JORGE, A. A.; LONDE, L. R.; REANI, R. T.; BACELAR, R. B.; SOKOLOV, I. M. Vulnerability analysis in complex networks under a flood risk reduction point of view. **Natural Hazards and Earth System Sciences Discussions**, v. 2019, p. 1–8, 2019. 3

TOMÁS, L. R.; SOARES, G. G.; JORGE, A. A.; MENDES, J. F.; FREITAS, V. L.; SANTOS, L. B. Flood risk map from hydrological and mobility data: a case study in São Paulo (Brazil). **Transactions in GIS**, v. 26, n. 5, p. 2341–2365, 2022. 3

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of 'small-world'networks. **Nature**, v. 393, n. 6684, p. 440–442, 1998. 1, 8

YANG, Z.; SONG, J.; GAO, S.; WANG, H.; DU, Y.; LIN, Q. Contact network analysis of Covid-19 in tourist areas——based on 333 confirmed cases in China. **Plos One**, v. 16, n. 12, p. e0261335, 2021. 1

YILDIRIMOGLU, M.; KIM, J. Identification of communities in urban mobility networks using multi-layer graphs of network traffic. **Transportation Research Part C: Emerging Technologies**, v. 89, p. 254–267, 2018. 1

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.