Modelling and Simulating Disruptions in Porto's Public Transports

Adriano Soares
Faculty of Engineering
University of Porto
Porto, Portugal
up201904873@edu.fe.up.pt

António Ribeiro
Faculty of Engineering
University of Porto
Porto, Portugal
up201906761@edu.fe.up.pt

Pedro Pinheiro
Faculty of Engineering
University of Porto
Porto, Portugal
up201906788@edu.fe.up.pt

***Abstract*—This study introduces a sophisticated model for Porto's public transportation, utilizing the STCP and Metro do Porto (MDP) networks, represented as a graph. With it we propose calculating the shortest paths between centroids of statistical sections from Census data, providing a nuanced view of the network's efficiency. Our analysis confirmed the critical role of the Trindade station and highlighted challenges in data consistency and quality. The discussions in our report focus on the implications of network modifications and the importance of coherent data sources in accurate modeling. Future enhancements will involve leveraging route data for decision-making, optimizing journey planning, and enhancing network resilience.**

***Index Terms*—Urban Mobility, Modeling, Simulating**

I. Introduction

As urban areas grapple with a surge in population growth, urban mobility is increasingly recognized as a pivotal concern due to its profound impacts on residents' daily lives and other aspects of the city, spanning the economy, environment, and overall quality of life. According to the most recent census data [1], it is estimated that 19,84% of the Portuguese population use public transport in their daily commutes, with an average travel time of 37,81 minutes, which is significantly higher than an individual alternative, at 17,81 minutes.

To better understand this area, urban planners often use modeling and simulation to apprehend the current state of the transport infrastructure and possible improvements. These tools must incorporate the complex aspects of transportation, such as stops, trips, number of passengers, schedules, and prices. There is a lack of a standardized mobility model for the city of Porto that integrates its two most significant public transport types (the bus and the metropolitan), which could be crucial in identifying infrastructure failures and efficiently collecting data for future decisions.

Our goal is to craft a descriptive and speculative model of Porto's public transportation network, aiming to authentically capture the nuanced behavior of public transport in real-life scenarios and explore hypothetical situations.

The remaining structure of this paper is as follows: Section II provides an overview and comparison of similar works done in this field, then Section III describes our datasets and analysis methodology, whose results are presented in Section IV. We end the document in Section VI, with a summary of the presented work and features we still want to incorporate in the future.

II. Related Work

Numerous works have focused on analyzing the connectivity of multi-modal public transport networks using different datasets, such as *GTFS*. Some deliver new graph theory approaches to network and disruption analysis, which are adaptable to multiple datasets, while others center around specific city-wide case studies.

In [2], the authors develop a graph-oriented analysis of *GTFS* and its improvement over standard transit indicators. They studied a time-expanded model of a small Montreal transit agency stored in a Neo4j database. In this graph, nodes represent a travel-related event (arrival, transfer, or departure), and edges represent relationships between both events, such as a connection that links a departure to arrival with the distance and duration. The authors include two levels of indicators: a stop-level indicator that describes the hourly connectivity of stops (and the extent of service offered) and another route-level that tackles the effectiveness of the speed in traveled road segments.

MUME [3] (Multi-modal Urban Mobility Estimation) models and assesses a multi-modal transport network as a multiplex network, stacking layers representing each type of transportation with intra and inter-layer edges. It approximates user behavior (traversing the graph) as Markovian processes and evaluates the robustness of the network with percolation theory [4], which can define how edge failure affects global connectivity. The paper also proposes a more efficient variation of [5], a numerical approach based on the eigendecomposition of the normalized supra-Laplacian, that estimates network coverage and case studies for Paris, New York, and London.

The city of Lisbon is studied in [6], which addresses common challenges in multi-modal descriptive analysis and enumerates different guidelines for the data consolidation process from a data analysis, decision level, and impact analysis perspective. The study leverages and consolidates known datasets from Lisbon's transport infrastructure, such as fare/validation collection, road junction detectors, and bike sharing data, integrating contextual information and its impacts into the model, namely the cultural agenda, commercial zones, education poles, sports buildings, health facilities, and road works. Different time granularities are considered: calendric (days of the week) and interval (peak hour intervals), and three depths of spatial referencing (TAZ, municipalities, and parishes). From the transport datasets, an OD matrix is compiled, with cross-carrier and cross-mode commutes labeling, from which they infer two index values that are modified to examine patterns of multi-modality (regarding the fraction of the population that uses a transport) and use intensity. One is the *Herfindahl–Hirschman Index* (HHI), which measures the pattern repetition of travel decisions and varies from 0 to 1 depending on the diversity and intensity of transport utilization (0 represents high variability and low individual intensity, while 1 represents a monopolization). The other is an adapted Gini coefficient [7], which expresses how habitual the activity-travel patterns of the population are.

One study [8] has an exclusive emphasis on Porto's mobility patterns, which crafts an *Urban Dynamic Indicator* (UDI) that aims to capture high spatiotemporal resolution of the city's pendular movements to help decision-making by the city planners and funds allocation. The indicator aggregates information from mobile phone locations, Waze traffic data, public transport, scooter utilization, and air quality. The researchers applied Factor Analysis [9] that, along with *Principal Component Analysis* (PCA), resulted in a composite value that retains underlying relationships across the variables, which are expressed as a linear combination. It provides insightful results of spatiotemporal variation of the UDI and areas of the city that need public transport improvement.

The key points from the related work are concisely captured and presented in the features outlined in Table I.

TABLE I: Overview of related work and their contributions.

Ref	**Required**					**Literature**	
	Porto's network	Node disruption	GTFS datasets	Multi-modal network	Fare/Ticket data	Edge disruption	Graph Theory metrics
[2]	✗	✗	✓	✓	✗	✗	✓
[3]	✗	✗	✓	✓	✗	✓	✓
[6]	✗	✗	✓	✓	✓	✗	✗
[8]	✓	✗	✓	✓	✓	✗	✗

The latest Porto mobility study does not incorporate a graph-based approach to infer its mobility indicator, measure network coverage, or simulate a disruption. Thus, we hope we can fill that gap with our contribution.

III. Methodological Approach

A. Problem

Due to the lack of availability of a consistent and updated model that can capture the complexity of the public transport coverage in the city's $41km^2$ area, our problem could be formulated as the following two research questions:

1) How can we create a platform that could provide an effective and comprehensive model that outputs meaningful data in a timely and effective manner that could support decision-makers?
2) How can we accomplish this efficiently in terms of computational resources, ensuring a timely and effective analysis?

B. Datasets

a) GTFS: The General Transit Feed Specification is a standardized format for public transportation information. In essence, it is a collection of text files that include stops, routes (and their shapes), stop times, trips, transfers, and calendars. For this study, we explored available datasets from the Open Data Repository of Porto[1] and used the 2022 collected data of both Metro do Porto and STCP.

b) Validations: This dataset integrates multiple *.xlsx* extended files with validation information for the bus and metro network from 2022 to 2023. Each row designates the number of validations that occurred in a particular timestamp and relevant information to specify the stop in which the collection was performed.

c) Geopackage (BGRI): The *geopackage* obtained from Instituto Nacional de Estatística (INE) encompasses demographic data collected in the 2021 Census [1] (e.g., a count of each age group). This format divides each of Porto's parishes into smaller geographical units, referred to as statistical sections, according to the *Base Reference Geographical Information* (BGRI), allowing us to obtain specific information on residents and calculate local centroids used in the network graph.

C. Graph Connectivity

After evaluating several modeling approaches and simulation-oriented software, a choice was made to represent the city and its public transportation network as a graph. The use of this simple yet powerful data structure meant that when compared to alternatives, for example, MATSim [10], further work would be required upfront, and fewer visual outputs would be easily achievable, while better performance and more customizability would be beneficial.

Graphs are well-suited for modeling mobility networks due to their ability to represent complex relationships between entities using nodes and edges. Nodes can signify locations, while edges depict connections such as roads. Graphs can also represent topology, directionality, and connectivity, with weighted edges allowing the incorporation of metrics such as distances, travel times, and routes. Lastly, graph algorithms to identify shortest paths have been extensively studied and optimized.

D. Graph Structure

Taking all the aforementioned advantages of modeling the network as a graph, we designed our public transport graph from the gathered dataset to fully represent bus and metro

[1] https://opendata.porto.digital/

travel. A segment of our network is pictured in Figure 1, in which the blue nodes represent stops, the green a geographical centroid, and the connections signify existing paths from one node to another. Our graph components are described as follows.

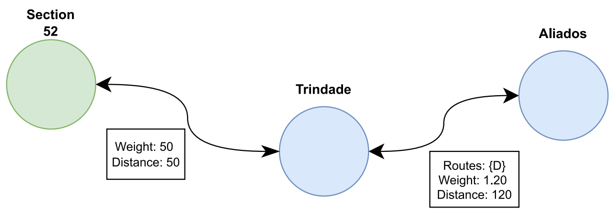

Fig. 1: Graph segment representation.

1) Nodes: Nodes in this work represent locations from where paths may begin, end, or go through. These can be centroids of statistical sections calculated from the locations of statistical sections obtained from the Geopackage. Alternatively, these can be public transport stops from either the subway (MDP) or the bus (SCTP). In both cases, nodes have accurate geographical locations and an accurate name or ID that enables their identification.

2) Edges: In this context, edges represent the voyage from one node to another. An edge between two centroids or a centroid and a stop represents walking from one location to the other, and edges between stops represent using public transportation to travel between the locations. While it is possible to walk from any point to another in the real world - doing so would leave us with a complete graph - the choice was to only create walking edges whose Euclidean distance would be under 2 kilometers. Considerably reducing the number of edges allowed for much better performance in further operations on the graph while still providing an accurate representation of it, as generally, individuals do not walk further than this distance while simultaneously using public transportation. Moreover, this optimization left us with a graph that is still strongly connected, a requirement for computing the shortest path between all centroid nodes.

Each edge has several attributes: the Euclidean distance between the nodes it connects, the weight of traversing such edge when computing the shortest path between two nodes, and the public transportation routes that go through them. The weight of the edge is the same as the Euclidean distance in the case of edges that represent walking. When edges represent traveling through the subway, the weight is 1% of this value. This is to account for the speed and commodity traveling that this means provides. Lastly, when using the bus, the weight is 5% of the Euclidean distance. To better simulate the distance traveled and considering that, unlike the subway, the bus's path is by no means straight, we took advantage of other data present on the GTFS unused until then. Specifically, we used positions in between each bus stop to better estimate the total distance said trip traveled, that is, the distance used to compute the weight of the edge.

E. Shortest Path Computation

Calculating the shortest paths between the centroids of statistical sections is a crucial aspect of our model, as it represents the optimal route for individuals traveling between centroids. The shorter the total path weight, the more efficient the route. Furthermore, by averaging the shortest paths from a centroid to all others, we can analyse the effectiveness of public transportation services of a given statistical section.

Various algorithms are available for computing shortest paths in graphs. Our graph, free of negative weights, is compatible with algorithms like Floyd-Warshall, Dijkstra's, and A*. The Floyd-Warshall algorithm initially seemed promising as it floods the graph to compute distances between every node. However, our graph is comprised of both centroid and stop nodes. Crucially, we only require distances between centroid nodes, not between stop nodes. This necessitated a solution to either filter paths post-calculation or modify the algorithm to exclude stop nodes from consideration. Upon investigation, we found such modifications to be impractical. Consequently, we initially opted for Dijkstra's algorithm, strategically avoiding nodes representing stops.

However, we encountered performance challenges. With 1.659 centroid nodes in our graph, the total number of paths to compute was the square of this number, 2,75 million shortest path calculations. The average execution time for calculating the shortest path using Dijkstra's algorithm was approximately 3 seconds. Based on this duration, the process would take around 3 months to complete. Seeking efficiency, we turned to the A* algorithm, using the Euclidean distance between centroids as a heuristic. This approach enhanced performance and yielded results consistent with those obtained through Dijkstra's algorithm, confirming its reliability. A comparison of the execution time of the shortest path calculus can be found in Appendix A, Figure 3.

F. Intel Dev Cloud

Despite the performance enhancements achieved through the A* algorithm, the computations were still demanding. An average calculation time of just one second, a rather optimistic figure given our machines' capabilities, implied a total duration of about 32 days for all calculations. This highlighted the need for greater computational power.

To address this, we turned to the free tier of Intel Dev Cloud[2], which provided access to multiple machines equipped with i9-11900KB processors and 32 GB of RAM. Utilizing a single machine, we reduced the execution time per calculation to 0,5 seconds, effectively halving the total computation time. However, this alone was not sufficient.

We further optimized our approach by adapting our script for threaded, parallel processing across multiple machines, dividing the workload efficiently. This adaptation also allowed for the script to be paused and resumed without data loss, a critical feature considering the 24-hour wall time limit per machine.

[2]https://devcloud.intel.com/oneapi/

Ultimately, these collective efforts significantly accelerated our process, concluding the entire task in just under 2,5 days with six machines running concurrently, a substantial improvement from the initial estimate of over a month.

G. Simulating disruptions

Up to this point, the discussed model was descriptive, which is useful as it gives urban planners the ability to assess the current state of mobility in the city. Modifying existing stop nodes seemed to be the best approach to further increase the extent to which this work can aid their decision-making. The creation of nodes would allow them to test the creation of stops, with the creation of several nodes being able to simulate the creation of an entirely new line. Modifying nodes is also an option, which could be used to simulate the relocation of a stop by altering its location. Lastly, the deletion of nodes simulates their unavailability, which could model situations in which a stop is unusable because of a natural disaster or maintenance. Since these approaches simulate different real-world scenarios but are similar in their implementation and challenges, we explore in this document the disruption of a stop, represented by removing a node.

The node we opted to remove was the subway stop of Trindade. This choice was intentionally made by considering the data we had collected during modelling. Specifically, Trindade was present in 1.331.816 shortest paths, which is over 48,4% of the total paths computed. The stop with the closest representation to this was Carolina Michaelis, which had less than half of the presence of Trindade. Having identified this outlier, which matched our perception and previous knowledge of Porto's subway network, we thought it interesting to simulate the effect the disruption of the city's main subway hub would have on the public transportation network overall.

After deleting Trindade's node from the graph, the need to re-compute the shortest paths between all centroids arose. To further optimize the computation for the particular scenarios where nodes are removed or modified, we analyzed the output of the shortest path's computations with the entire graph and identified the paths the modified or removed node was part of - the 48,4% of the paths previously mentioned. Then, instead of calculating the shortest path between all centroids, only the pairs whose path included the node were computed, reducing the computational time significantly.

H. Incorporating Validations

To enhance our model's accuracy and comprehensiveness and enable more extensive measurements, we integrated the latest dataset featuring *.xlsx* files. These files provide detailed validations per stop per hour, enriching our model with crucial data. The primary objective of this integration was to gauge the attractive power of statistical sections more accurately. Validations effectively measure a statistical section's relative allure, as stops with higher validation frequencies typically indicate popular gathering points, such as malls or hospitals.

Typically, individuals using public transportation validate their tickets at stops within their residing statistical section. They may also do so at interchange stops en route to their destination. Ultimately, they validate tickets at their final destinations, which accumulate a higher number of validations due to their attractive nature. This phenomenon is because these locations draw individuals from various statistical sections, who validate their tickets upon returning home. A heatmap vividly demonstrates the validation distribution in both simulation scenarios, with and without the Trindade stop, as shown in Figure 5.

Our approach involved aggregating the total validations for each stop from the dataset and associating them with corresponding nodes in the graph. Integrating these validations into our model was feasible thanks to certain assumptions, though not without challenges. A primary challenge was the mismatch in stop names between the validations dataset and the *GTFS* files. We addressed this by employing *difflib*, a Python package, to match stop names from both datasets based on the similarity of their character sequences.

We observed that in the bus *GTFS* files, some stops shared the same name but had different identification codes. For instance, "Montes Burgos" appeared as MTB1 and MTB2, varying with the route direction. In such 1-many relationships, we divided the total validations evenly among the stops (e.g., the collective validations for "Montes Burgos" were split between MTB1 and MTB2). This approach more accurately reflects reality than duplicating validation counts across all stops.

Lastly, we noted some stops in the *GTFS* were absent in the validations dataset. For these, we assigned a number of validations equal to the average of validations across all stops. This method ensured a realistic representation of stops lacking specific validation data, maintaining the integrity of our model's overall analysis.

IV. Results

Our analysis can be broken down into a numerical analysis of multiple metrics that we gathered from the final dataset and two maps that provide a visual representation that better expresses our results.

We found that the Pearson correlation between the number of validations and the number of stop appearances in the shortest paths is 0,69. This metric (defined in Figure 2) represents how two series of points are related to one another, where 1 symbolizes a perfect linear relationship (a proportional increase when the other increases). We'll discuss the meaning and implications of this value in the next Section V.

$$\rho = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Fig. 2: The Pearson correlation coefficient (ρ) formula using covariance ($\mathrm{cov}(X,Y)$) and standard deviations (σ_X and σ_Y).

Our numerical analysis is depicted in Table II. We observed a decreased average distance between statistical sections when Trindade was removed.

TABLE II: Analytical network results.

Metric	w/ Trindade	w/o Trindade
Centroid-centroid Dist.	8158,96 ± 4488,98 m	7856,51 ± 3907,33 m
Avg. weight	622,06 ± 33381,66	675,41 ± 44652,96

Our obtained results are plotted in Figure 6 at Appendix B, with two map representations of the average weight (mentioned in Section III-E), when travelling to a geographical centroid. This metric conveys how difficult (time-costly) it is to travel to a specific region of Porto's municipality, averaging the weight of all possible shortest paths from all starting centroids in the rest of the map. The higher the weight value is, the warmer the colour each section acquires.

V. Discussion

The high correlation found between the number of validations and the number of stop appearances in our data validates our approach with real-world observations. An increase in the number of validations in a given stop represents its usefulness to users and by intuition that is justified by its inclusion in more sets of shortest paths, given that a user will choose a stop that is quicker to reach a destination, validating the ticket in that station and not elsewhere.

The distance required to travel from one statistical section to another typically spans 8158,960 ± 4488,982 meters. However, this figure reduces to 7856,508 ± 3907,331 meters following the removal of Trindade. Accompanying this change, we observed an 8% increase in the average computed weight, escalating from 622,058 ± 33381,66 to 675,41 ± 44652,96. The slight uptick in the average weight post-removal of Trindade underscores its vital role in the network, serving as a key conduit in numerous shortest paths and essential for maintaining an efficient schedule.

The decrease in the pure distance value is a direct consequence of how the algorithm recalculates the shortest paths in Trindade's absence. As Trindade serves as a central node, its unavailability necessitates the consideration of bus transport as an alternate route. Bus routes make more frequent stops, thereby creating a denser stop distribution across each map section. This increased density enables users to disembark closer to their final destinations or centroids. Consequently, this shift leads to the generation of shorter paths which, although potentially more time-consuming, effectively reduce the overall distance travelled, illustrating a nuanced interplay between path length, travel time, and network efficacy in urban transportation planning.

Finally, the two maps in Figure 6 at Appendix B provide a clear look at how the weight distribution disperses across the map. We find in both graphics that the centre region of Porto has a much darker presentation. This means that travelling to and between those statistical sections is much less time-intensive than travelling to the municipality's periphery. This can be justified by real-world examination: the centre region of Porto has more bus stops and routes available in city population hotspots, which are closer to one another and capacitate users to exchange stops on foot more easily and transfer between routes. The metro stations in this area are also closer and underground, which decreases travel time.

Travelling to the periphery, which is presumably taking public transport home (although the validations dataset is not detailed enough to substantiate this point), is slower, not only because the distance increases but also because the network is less dense. The second map reiterates the fact that Trindade is a critical stop, considering not only the faster speed reached by the metro but also its well-connected nature. Despite having a 10% increase noted in the colour bar, we find that the overall coverage pattern remains relatively the same.

VI. Conclusion and Future Work

In this study, we successfully met our initial goal of developing a model capable of accurately representing Porto's transport infrastructure, which is both adaptable and capable of incorporating additional relevant data. For instance, the integration of parking lot locations as network nodes in our model, with corresponding contextual values for the edges, could significantly enhance the model's utility. This enhancement would not only improve the management of the parking market but also open up new avenues for transportation usage and opportunities.

Our simulations affirm the pivotal role of the Trindade station within the metro network, identifying it as a critical junction essential for maintaining optimal schedules and connections. This finding underscores the need for robust resilience measures to safeguard against potential disruptions.

A key learning from our endeavour is the importance of high-quality datasets and the complexities posed by inconsistent data sources. These challenges necessitated the adoption of heuristic approaches and approximations in our methodology. Additionally, establishing efficient workflow pipelines was crucial in managing large volumes of data and driving innovations in code efficiency.

Looking ahead, there is scope for further enhancements to our platform. One potential area of expansion involves leveraging the minimal set of routes encompassed in each shortest path calculation. This information could be instrumental in informing decision-making processes, both from the perspectives of users and planners. It opens up possibilities for selecting the most efficient routes for travel between any two points and optimizing journey planning in urban transport networks. An insight into initial work on this area can be found in Appendix C, Figure 7.

The project's code and workflow pipeline are freely available in a GitHub repository[3].

References

[1] I. N. de Estatística, "Censos 2021. xvi recenseamento geral da população. vi recenseamento geral da habitação : Resultados definitivos." Lisboa : INE, 2022. Disponível na https://www.ine.pt/xurl/pub/65586079.

[3]https://github.com/francisco-rente/MS-2023

[2] P. Fortin, C. Morency, and M. Trépanier, "Innovative gtfs data application for transit network analysis using a graph-oriented method," *Journal of Public Transportation*, vol. 19, no. 4, pp. 18–37, 2016.

[3] A. Baggag, S. Abbar, T. Zanouda, and J. Srivastava, "Resilience analytics: coverage and robustness in multi-modal transportation networks," *EPJ Data Science*, vol. 7, pp. 1–21, 2018.

[4] R. Albert, H. Jeong, and A.-L. Barabási, "Error and attack tolerance of complex networks," *Nature*, vol. 406, p. 378–382, July 2000.

[5] M. D. Domenico, A. Solé-Ribalta, S. Gómez, and A. Arenas, "Navigability of interconnected networks under random failures," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8351–8356, 2014.

[6] C. Lemonde, E. Arsenio, and R. Henriques, "Public transportation multimodality in the city of lisbon," *Transportation Research Procedia*, vol. 58, pp. 75–82, 2021. XIV Conference on Transport Engineering, CIT2021.

[7] M. Diana and M. Pirra, "A comparative assessment of synthetic indices to measure multimodality behaviours," *Transportmetrica A Transport Science*, vol. 12, no. 9, pp. 771–793, 2016.

[8] B. Jardim, M. de Castro Neto, and P. Calçada, "Urban dynamic in high spatiotemporal resolution: The case study of porto," *Sustainable Cities and Society*, vol. 98, p. 104867, 2023.

[9] H. F. Kaiser, "The application of electronic computers to factor analysis," *Educational and psychological measurement*, vol. 20, no. 1, pp. 141–151, 1960.

[10] A. Horni, K. Nagel, and K. W. Axhausen, eds., *The MATSim Book*. London: Ubiquity Press, 2016.

Appendix A
Algorithm Execution Times

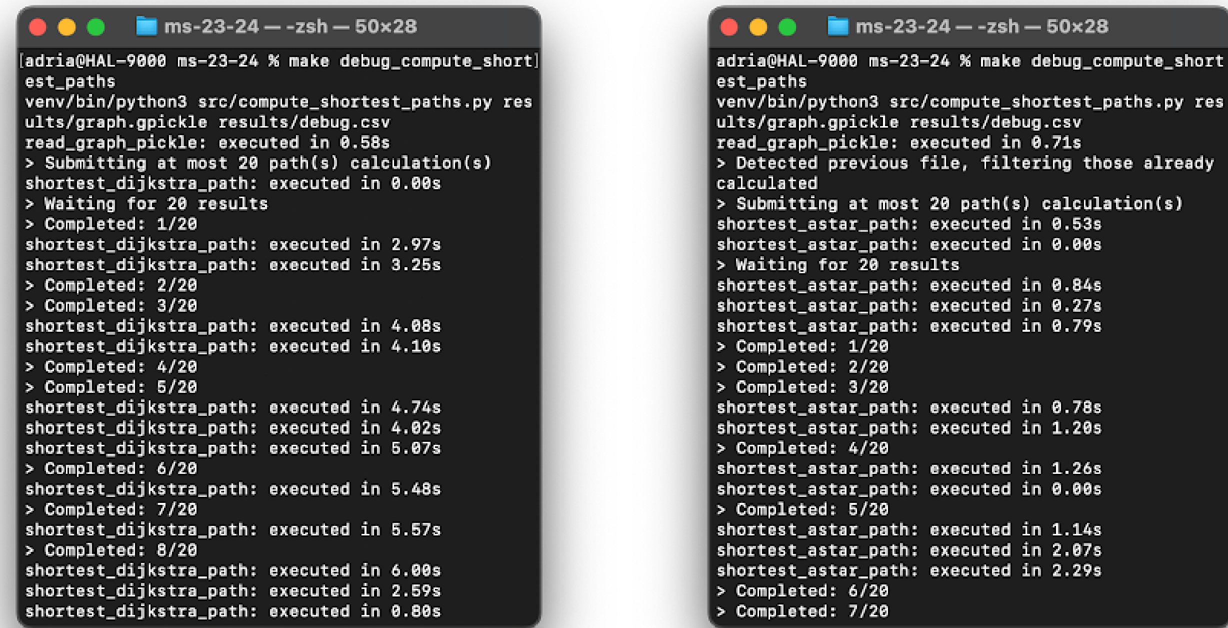

Fig. 3: Comparison of Dijkstra (left) and A* (right) algorithms execution time.

APPENDIX B
RESULTS

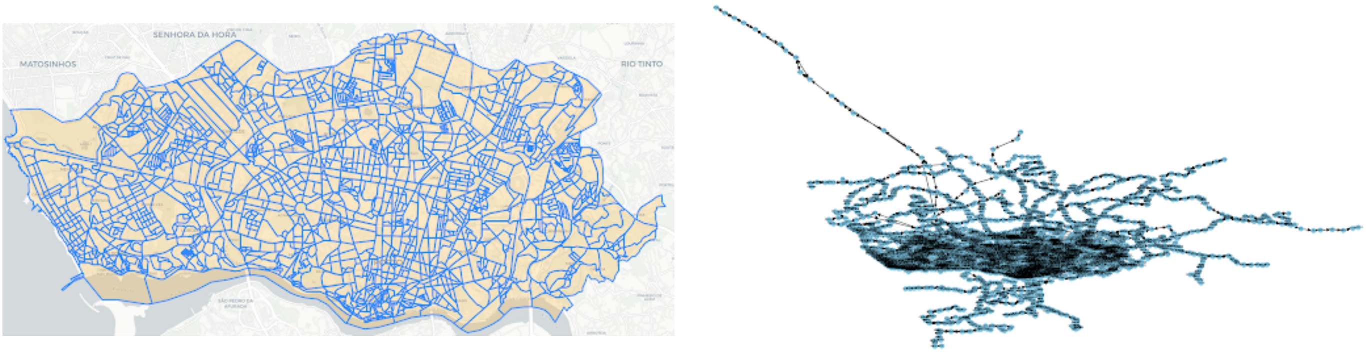

Fig. 4: Statistical sections of the city of Porto (left) and MDP and STCP routes of the Porto's district (right).

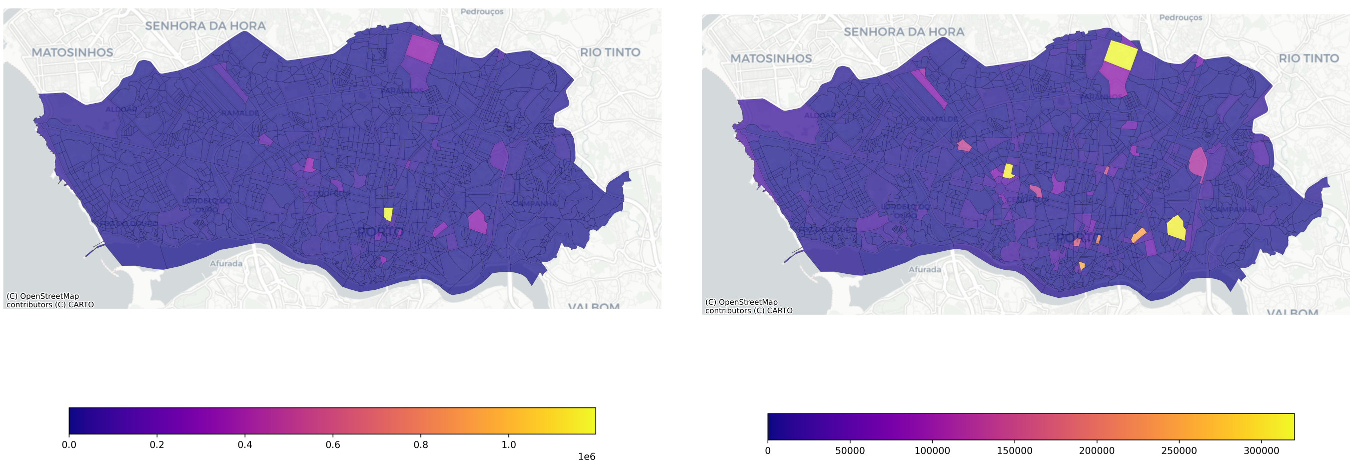

Fig. 5: Collected user validations for the PT network with Trindade (left) and without (right).

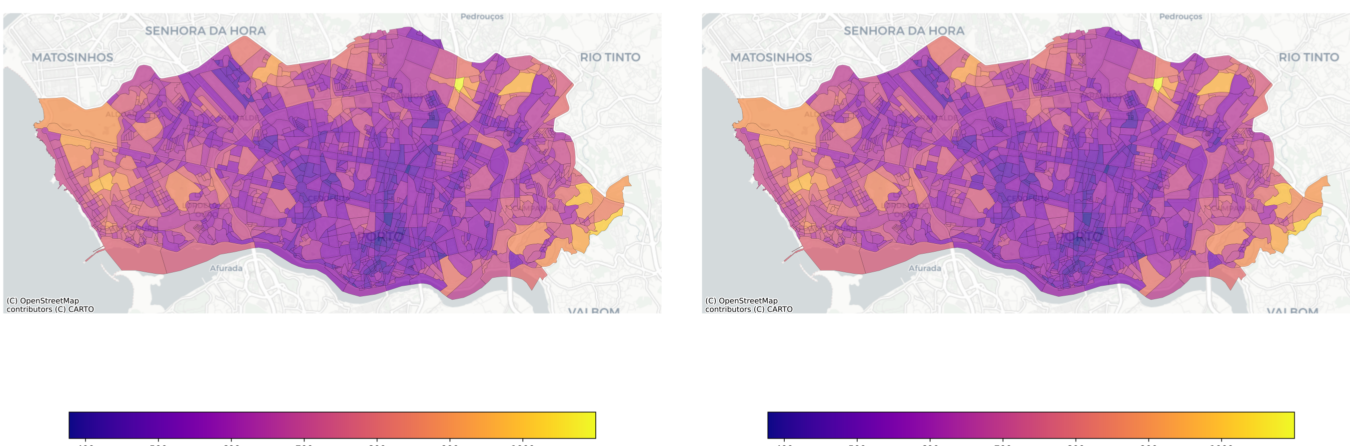

Fig. 6: Computed travel weight for the PT network with Trindade (left) and without (right).

APPENDIX C
FUTURE WORK

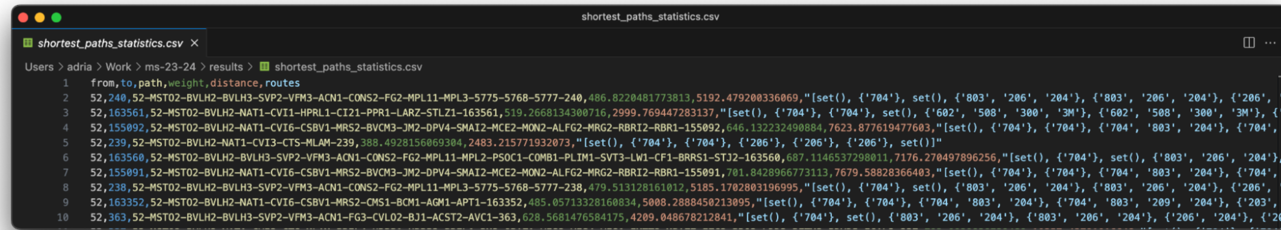

Fig. 7: Shortest paths from any statistical section to another with route information.