Epiflow: a hybrid approach to track infectious disease spread in Brazil based on travel data and graph databases

Mariama C. S. de Oliveira[1], Andrêza Leite de Alencar[2], Natalia Tatiele S. de Oliveira[1], Lucas Henrique Gonzaga de Sales[2], Antônio Ricardo Khouri Cunha[3], Pablo Ivan Pereira Ramos[3]

[1]Centro de Informática – Universidade Federal de Pernambuco (UFPE)

[2]Departamento de Computação – Universidade Federal Rural de Pernambuco (UFRPE)

[3]Instituto Gonçalo Moniz – Fundação Oswaldo Cruz
mcso@cin.ufpe.br, andreza.leite@ufrpe.br, ntso@cin.ufpe.br,
lucas.gonzaga@ufrpe.br, ricardo.khouri@fiocruz.br, pablo.ramos@fiocruz.br

Abstract. *Based on open data on cities and transport, the present study proposes an approach that uses travel probabilities and graph-oriented database to identify possible disease propagation routes within the Brazilian territory. Route identification was implemented by adapting the Dijkstra algorithm in the Data Science module of Neo4j. A tool called Epiflow was also developed to allow visual exploration of the proposed approach. Validated by COVID-19 data, the approach successfully predicted routes for large geographical areas of risk, such as states. These findings suggest that transport data and graph databases can be used to create applications that assist decision-making in tracking disease spread in the early stages.*

1. Introduction

According to World Health Organization (WHO), until November of 2022, more than six million people had died due to COVID-19 [WHO 2022a]. It is hard to ignore the dramatic effects that the pandemic has brought to the world. The problems impact diverse areas, including health, education, and the economy. In short, the fallout of an epidemic disease is tremendous, influencing virtually every sector and population.

Another factor that heightens the likelihood of new pandemics arising is human mobility [Mu et al. 2021, Bajardi et al. 2011, Peixoto et al. 2020]. Today, the world is highly connected by various means of transportation. In 2019, about 38.9 million flights were performed in the world [Statista 2022]. In Brazil, about 1.39 million interstate bus trips were performed in 2019 [Ministério da Infraestrutura 2020]. In the face of this potential hazard, recent studies address the movement of humans, and its consequences on spreading infectious diseases [Mu et al. 2021, Bajardi et al. 2011, Peixoto et al. 2020]. Such studies are aided by the massive amount of data available today, which makes following human displacements easier.

Based on the context of epidemics and human mobility, the present study aimed to explore an approach using city and transport data arranged in a graph structure to examine how infectious diseases spread in Brazil. The study culminated in elaborating a visualization tool called Epiflow, which allows users to explore potential diseases spreading

throughout the country. In order to describe and discuss this process, the paper is organized into six sections. The first section, Introduction, defines the scope of the study and its significance, along with its objectives. The second section, Literature Review, presents the prevalent techniques for tracking disease spread. The third section, Methodology, describes the data, computation techniques used, and the developed application. The fourth section, Evaluation and Results, covers the validation process and its outcomes. The fifth section, Discussion, discusses the paper's discoveries, limitations, and future works.

1.1. Background and significance

To prevent other pandemics from arising, a domain called Genomic surveillance emerged. In tandem with traditional epidemiological approaches, Genomic surveillance aims to monitor pathogens continually and analyze their genomic similarity and disparities [WHO 2022b]. In March 2022, WHO released a ten-year strategy [WHO 2022b] to increase initiatives around the globe related to Genomic surveillance. According to the agency, COVID-19 has brought to light the implication of such actions by showing the importance of tackling epidemic risk at early stages. Accordingly, an initiative named ÆSOP[1] (Alert-Early System of Outbreaks with Pandemic Potential) was formed in Brazil. ÆSOP is a data-driven system that hopes to alert the country at the early stages of potential respiratory viral disease outbreaks. According to them, one of the biggest challenges the initiative faces is defining the best strategy for sampling.

In this context, the present study presents itself as an opportunity to explore and understand the propagation behavior of infectious diseases using human mobility data and city connections in a graph structure. So that strategies for sampling and other public policies for addressing disease outbreaks at early stages can be defined.

1.2. Aim and Objectives

The present study aims to develop an approach based on cities and transport data to identify possible routes an infectious disease can take during its spreading process. It expects to analyze the propagation behavior in the Brazilian territory, utilizing graph structure and travel probability between cities.

To accomplish the main goal, the study posed three specific objectives.

1. To identify useful datasets and perform the ETL process on the chosen data (city, health service, and transportation data);
2. To propose a solution that recommends and ranks which cities should be investigated in case of finding evidence of infectious disease in a particular city;
3. To develop a system capable of identifying propagation routes to specific cities.

2. Literature Review

The purpose of this section is to present studies that address spatial disease spreading, primarily from the perspective of human mobility.

Many studies [Mu et al. 2021, Bajardi et al. 2011, Peixoto et al. 2020] indicate human mobility as a significant factor in the spatial spread of infectious diseases. As a result, we have found a considerable amount of studies employing models of spacial

[1] ÆSOP: `http://aesop.health/about-us`

transmission, in particular, the metapopulation model. Metapopulation is one of the simplest models of spatial modeling [Keeling and Rohani 2008]. The basic idea behind the model is to divide the population into subpopulations that have their own internal dynamics and eventually interact with each other. Usually, each city is considered a subpopulation, and the flow of people between them is the interaction. Even though this model is widely used, its downside is the need to find the proper division of a subpopulation since the model assumes that each subpopulation is homogeneous. In addition, it is important to estimate the flow between connections correctly [Balcan et al. 2010].

Another possible approach is the Agent-based model. In this model, each individual is considered an agent interacting with another. Based on the interactions between agents, it is possible to define the behavior of the disease being transmitted. During the COVID-19 pandemic, Wei et al. [Wei et al. 2021] implemented an intercity multi-agent model. According to the authors, the model could estimate early infections in China with high precision. However, it is important to highlight that this approach is computationally costly once it requires tracing the behavior of each individual. This might be impracticable when the population observed increases on a global scale.

Effective Distance is another spatial model worth mentioning. The concept introduced by Brockmann et al. [Brockmann and Helbing 2013] states that it is possible to calculate an effective distance that represents the connectedness between cities. In other words, cities with smaller effective distances are more connected and more likely to propagate infectious diseases to one another. Based on this idea, Saderkar et al. [Sadekar et al. 2021] created an infectious diseases hazard map in India.

The spatial models presented until now (except for Effective Distance) incorporate classical epidemiological models into their computations, which consider disease behavior in their estimation. However, there are simpler approaches, such as calculating the flow probability between cities [EpiRisk 2022, Gilbert et al. 2020, Nakamura and Managi 2020]. The work of Gilbert et al. [Gilbert et al. 2020], for instance, calculated the vulnerability of African countries during the COVID-19 pandemic. Unlike other approaches, such models do not require a deep knowledge of disease-spreading behavior and are computationally less expensive.

In light of the previous studies, the present work focused solely on human mobility between cities rather than disease characteristics to reach its objectives. The solution will use a computationally non-costly approach like the one introduced by Gilbert et al. [Gilbert et al. 2020]. Similarly to their work, we will use traveling probabilities to track disease-spreading spatial behavior. This will allow us to create a tool that provides quick feedback to users, helping them explore different aspects of the problem. And unlike previous studies, which majorly concentrate on air transportation at a global scale, we consider land and air transportation at a Brazilian level. As a result, our analysis should provide a more accurate representation of disease spread in Brazil.

3. Data and Methodology

In this section, we will discuss the data used in the study and the ETL process that was performed on it. Furthermore, we will outline the tasks that were conducted to achieve the objectives of the study. Finally, a brief explanation of the visualization tool called Epiflow will be provided.

3.1. ETL process and Data

To implement the algorithms and the application described in the upcoming sections, an ETL (extract, transform, load) process (Figure 1) was performed on the data. Initially, the data was extracted from various governmental agency repositories. Subsequently, it underwent a series of transformations, including cleaning, integration, and estimation of missing data. Finally, in the load phase, the data was stored in the chosen databases.

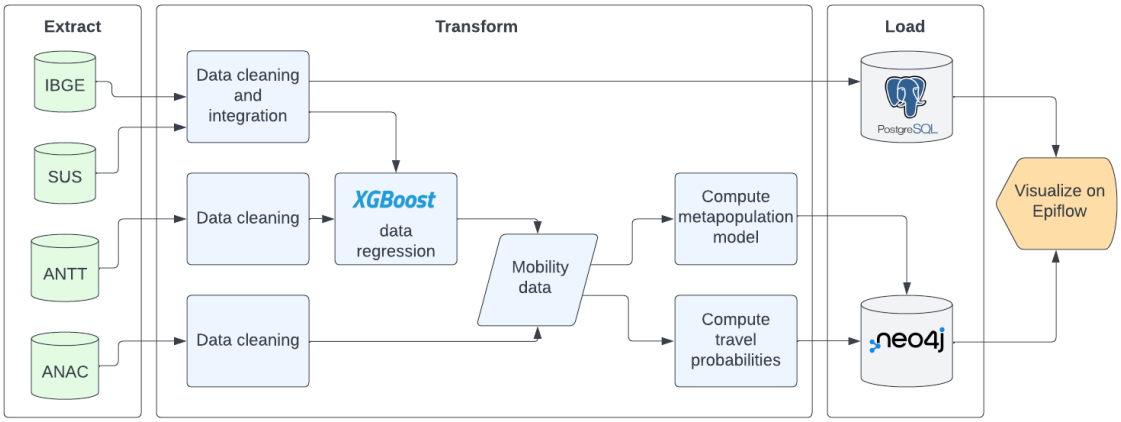


Figure 1. Flow diagram of the ETL methodology adopted.

The present study utilized the data provided by Brazilian governmental institutions, the Brazilian Institute of Geography and Statistics (IBGE[2]), Unified Health System (SUS[3]), and transport regulatory agencies - National Civil Aviation Agency of Brazil (ANAC[4]) and National Land Transportation Agency (ANTT[5]). Unfortunately, most of the data supplied by these agencies were not ready for use. Therefore, it had to undergo a transformation phase, primarily with Pandas [Reback et al. 2020], so we could utilize them. The subsequent paragraphs will present the data characteristics and the transformations performed. To learn more about the data, please consult the project data dictionary[6].

City data

Almost all city data was obtained from IBGE. The institute provided vital information such as estimated population, gross domestic product (GDP), number of hospital beds per inhabitant, and level of influence of Brazilian cities. In total, the city data amounts to 5570 municipalities with 31 features.

Health-related data

The health data was acquired from the agencies IBGE and SUS. Three types of data were collected from these sources, the flow of patients in the Brazilian territory (27750 rows),

[2] IBGE: `https://bit.ly/regic_ibge`
`https://bit.ly/area_municipality_ibge`
`https://bit.ly/population_ibge`
`https://bit.ly/api_ibge`

[3] SUS: `https://bit.ly/health_region_sus`
`https://bit.ly/reference_hospitals`

[4] ANAC: `https://bit.ly/data_anac`

[5] ANTT: `https://bit.ly/data_antt`

[6] Epiflow data dictionary: `https://bit.ly/epiflow_data_dictionary`

the territorial division of health regions (450 rows), and reference hospitals[7] (262 rows). This information enabled us to explore the flow of ill people in the country and consequently track the propagation of emerging infectious diseases as described in subsection 3.3.

Transport data

The transportation data refers to the number of airplane and bus passengers in 2019 within the Brazilian territory. They were obtained from the Brazilian transport regulatory agencies, ANAC and ANTT. As these are the primary means of transportation used for long trips in Brazil [Ministério da Infraestrutura 2020], it is possible to affirm that this data encompass a significant portion of passengers who travel in the country. However, during the analysis of the data, it was found that the number of passengers reported between certain city pairs needed to be adjusted. This was due to the fact that ANTT does not require bus companies to gather data within the same state, resulting in an underestimation of these records. In order to bypass this issue, the study utilized an XGBoost regressor[8] to estimate some trips. This resulted in estimating all bus trips within the same state and some interstate journeys; note that this issue did not affect airplane data.

3.2. Databases and data modeling

Both relational and non-relational databases, PostgreSQL [PostgreSQL 2022] and Neo4j [Neo4j 2022], were employed to develop Epiflow. PostgreSQL was able to load general information quickly in the application, while Neo4j was responsible for modeling relationships between cities. The latter was chosen due to its native graph storage, processing, and vast open-source data science library [Neo4j GDS 2022]. In this work, we will focus on Neo4j (graph). Figure 2 shows the Neo4j database schema utilized in the study.

We utilized a directed graph to represent the Brazilian cities network. In the graph structure, each city was represented as a vertex, and the edges connecting them represented the relationships between cities (see Figure 3(a)). Besides the Brazilian cities, the graph structure also includes other nodes that store information about the studied network, as shown in the graph schema in Figure 2; nevertheless, we will focus on the relationship between cities because they are the ones that provide the information needed to identify cities at risk and calculate propagation routes in the application. As the graph is directed, the relationship between cities has a source and destination. As we can see in Figures 2 and 3, the relationship between cities included two types of edges: `TRANSPORT_FLOW` and `HEALTH_FLOW`. `TRANSPORT_FLOW` comprises data related to bus and air transportation. This relationship has four attributes: air, bus and total (air + bus) flow in

[7] The hospitals of this network are also called Sentinel Service

[8] The XGBoost regressor was chosen due to its high performance in a variety of problems and robustness to outliers. The model utilized 47 features (no feature selection was applied), which included mainly the city and city connection attributes found in the data provided by IBGE. During the learning process, it was utilized 19378 observations that were divided into the train (75%) and test (25%) sets. It was utilized a k-fold cross-validation [Berrar 2019] with five folds while tuning the model and ten folds to obtain the final model's MSE metrics for the train set. The MSE and R^2 metrics for the test set were also calculated. The model's metrics are as follows: MSE mean (10 k-fold) = 0.407; MSE std = 0.024; MSE (test set) = 0.409; R^2 (test set) = 0.617.

probability terms, and the total number of individuals traveling in this connection annually. On the other hand, `HEALTH_FLOW` represents the flow of patients between cities (probability of a patient seeking a health service in another city) and has two attributes: high-complexity health service flow (high-cost treatments that usually involve hospitalization) and low and middle-complexity health service flow (medical appointments and exams, minor surgeries etc). The attributes regarding flow were represented in probability terms as it is more appropriate to estimate the risk of disease propagation. Section 3.3 will explain how these probabilities were obtained.

Figure 2. Graph schema of the database utilized in the study. The colored circles in the schema represent the entities modeled in the Neo4j database. `City`, `Municipality`, `State`, `PopulationArrangement`, and `Hierarchy` represent information regarding the official territorial division of Brazil, while `HealthRegion` and `ReferenceHospital` store information regarding the Brazilian health system. The arrows in the schema indicate how these entities relate to one another.

(a) (b)

Figure 3. Examples of the Neo4j graph structure. (a) The image illustrates two Brazilian cities connected by two types of edges, transport and health service flow. (b) The network structure of all cities connected to Manaus-AM.

3.3. Identifying cities at risk

Human mobility plays a vital role in the spread of infectious diseases, as previously discussed. With this premise as a starting point, we define cities at risk as the ones with the highest probability of receiving people from cities where evidence of infectious diseases wore found. Other projects [EpiRisk 2022, Gilbert et al. 2020] have inspired this approach since it is a not computationally expensive manner of calculating risk areas of pathogens propagation. Computing this probability is simple. It is only necessary to know how many people travel from one city to another, as we can see from Equation 1. Since this data was obtained from the transport agencies and regression, we were able to calculate the probability of traveling between cities. This probability also represents the risk a given city may face if it finds evidence of pathogens in one of its connections.

$$Pr(\text{travel from city A to city B}) = \frac{\text{Number of passengers from city A to city B}}{\text{Total number of passengers from city A}} \quad (1)$$

While the computation presented in Eq. 1 was necessary to determine traveling probabilities based on passenger flow, the flow of patients was obtained directly from IBGE, which had statistics on the subject. We consider that both flows are essential to identify cities at risk. However, explorations of propagation routes relied solely on passenger flow, as the next section will discuss further.

3.4. Identifying propagation routes

The purpose of this section is to explain how the probabilities of passenger flow stored at the edges of the graph were used to determine the most probable propagation route to a certain city or set of cities.

The edges of the graph store the traveling probabilities. Consequently, the most likely spreading path between any two cities is the one that has the highest likelihood of occurring. It is possible to compute it using Dijkstra's algorithm [Dijkstra 2022]. Originally, the algorithm was designed to find the shortest path between two nodes in a weighted graph. For example, it is useful for finding the quickest route between two cities. Three factors are crucial to determining the shortest path: the costs, the total cost, and the function used to compute it. The cost can be translated as the weight of each edge, the total cost is the sum of all edge weights that comprise the path (initially, it is 0 and increases as we walk through the graph's edges), and the function used is the sum. However, over the years, algorithm variants have been proposed to tackle different problems. Finding the most probable path, for instance, is one of them. These variants tinker with some algorithm characteristics like costs and function. Due to this, the version used in this work was slightly modified to find the most likely path. The modifications were the following:

1. A multiplication function was used instead of a sum function to calculate the total cost (since we are working with probabilities values).
2. The initial cost, instead of being 0, was -1 (to prevent the calculations from being 0 and force negative results);

Figure 4 illustrates how the algorithm with these changes works. If someone in city A decides to go to city C, the most probable path will be $A \rightarrow B \rightarrow C$ instead of $A \rightarrow C$ because the probabilities are $0.5 \cdot 0.5 = 0.25$ and 0.2, respectively. However, since Dijkstra chooses the shortest path, it is necessary to change the initial cost to -1, so the answer to the problem is $-1 \cdot 0.5 \cdot 0.5 = -0.25$ and $-1 \cdot 0.2 = -0.2$. In the end, -0.25 is the shortest path as well as the maximum probability as a negative value.

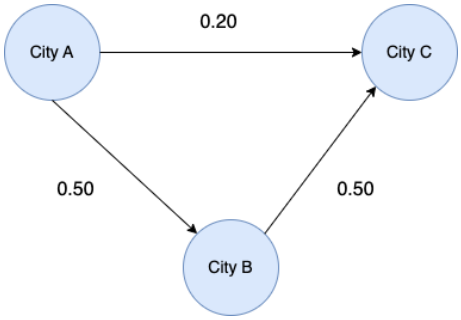
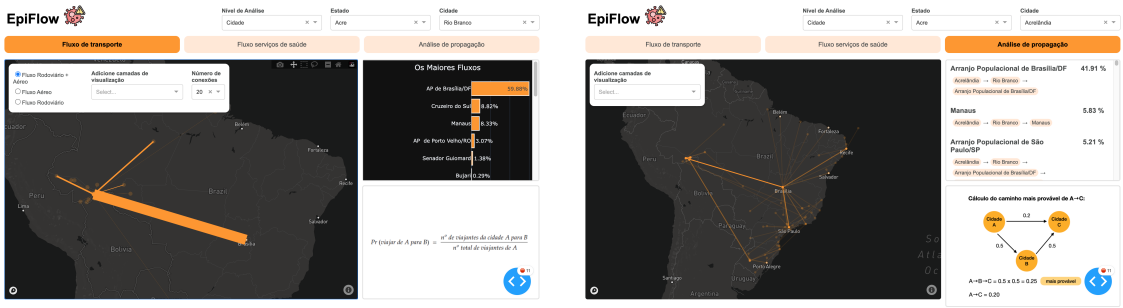


Figure 4. The directed graph shows three connected cities with their traveling probabilities.

The study utilized the built-in Neo4j's Dijkstra implementation [Neo4j GDS 2022]. However, it was necessary to modify the library's open-source source code [9] to add the mentioned changes.

3.5. Epiflow application

To showcase the study's objectives, the team created a Dash [Dash 2022] application named Epiflow. This interactive visualization tool allows the user to visualize the results of the tasks described in the previous sections. Therefore, it serves as a tool for exploring different disease-spreading scenarios. The application has two main functionalities: (1) To visualize the flows (transport and health service) originating from a selected city (Figure 5(a); (2) To trace possible spreading routes from a selected city to other cities (Figure 5(b)).



(a) (b)

Figure 5. Epiflow screens: (a) Visualization of the transport flow. (b) Visualization of spreading routes to specific cities.

[9] Neo4j Graph Data Science - Github: `https://github.com/neo4j/graph-data-science`

The source code and datasets used in the solution are available in a code repository[10].

4. Evaluation and Results

After implementing these tasks described in section 3, we evaluated them with data series from two COVID-19 variants highly spread in Brazil, alpha and gamma. Alpha was the first variant introduced in Brazilian territory. Most of the data indicates that the city of São Paulo was the country's entry point for the virus. That is an assumption that we will hold during this evaluation. As for the gamma variant, we will consider its epicenter the city of Manaus since, according to genetic surveillance [Faria et al. 2021], this variant emerged in this region.

4.1. Cities at risk

We evaluated the system's ability to identify cities at risk using the Spearman rank correlation metric. This metric helped us measure the relationship between the cities suggested by the system and the actual data. Initially, we ranked each region based on the number of COVID cases reported outside the epicenter, using a ranking system to measure which Brazilian regions the COVID variant reached first. Moving average thresholds were used to determine whether a region had enough cases to enter the ranking list. Ranks will lower as a region takes longer to reach the threshold. Once all regions were ranked, we calculated the correlation between the ranks and the probability provided by the system. Since there was insufficient data available at the city level for both variants, we assessed the system's effectiveness by analyzing the disease's spread on a state level. This adjustment was necessary to evaluate the system's proposed approach. The results for both epicenters can be found in Table 1.

Table 1. Spearman's correlation between the system's probability of locations at risk vs. the actual spread of COVID-19

	Moving average threshold		
Variant	0.5	1.0	1.5
Alpha	r(24) = -.87, p = .000	r(24) = -.82, p = .000	r(24) = -.84, p = .000
Gamma	r(18) = -.38, p = .110	r(18) = -.20, p = .410	r(18) = -.19, p = .440

According to Table 1, all the computed values for the variable alpha show a strong correlation between the locations at risk identified by the system (in our case, states at risk) and the ranked list of states. However, no statistically significant values were found when validating with the gamma variable since $p > .05$, so no inferences could be made.

4.2. Propagation route

In order to validate the most probable route, we use the same approach as when validating cities at risk (see subsection 4.1). However, in this study, we will only verify the probability of disease spreading, not the chosen path per si. Hence, we cannot determine if the suggested path is the most probable, but we can evaluate it in terms of disease-spreading probability. Once again, the considered epicenters were São Paulo and Manaus. We calculated the most probable routes from these cities to all other state capitals in Brazil. Then

[10] Code repository: `https://github.com/mariamaOlive/alerta-pandemia`

the study employed Spearman's rank correlation to determine the correlation between the ranking of capital cities[11] (explained in subsection 4.1) and the likelihood of a disease spreading along a particular route. The results are found in Table 2.

Table 2. Spearman's correlation between the system's path spread probability vs. the actual spread of COVID-19

	Moving average threshold		
Variant	0.5	1.0	1.5
Alpha	r(24) = -.82, p = .000	r(24) = -.83, p = .000	r(24) = -.81, p = .000
Gamma	r(24) = -.40, p = .042	r(24) = -.14, p = .496	r(24) = -.11, p = .584

In Table 2, there is a strong correlation between the likelihood of disease transmission through a particular route and the actual data of the alpha variant. Additionally, there is a moderate correlation ($r(24) = -.40, p = .042$) between the system's reported probability of disease transmission and the gamma variant spreading data. However, this inference only applies to the threshold of 0.5, the only statistically significant value for the gamma variant ($p < .05$).

5. Discussion

Based on the collated results, it is possible to draw some interesting conclusions regarding identifying cities at risk and propagation routes of disease propagation within the Brazilian territory. This section will reflect upon these findings, as well as the research process. Furthermore, the work's implications and limitations will also be discussed. Finally, this section will conclude by presenting ideas for future works.

The evaluation results show a clear correlation between the system's travel probabilities and the order in which the alpha variant spread to different Brazilian states. However, the same correlation could not be established for the gamma variant data due to not statistically significant results. Regarding the propagation route, there is a strong and moderate correlation between the system's estimated probabilities and the order in which the alpha and gamma variants first spread to the Brazilian capitals. Various moving average thresholds (0.5, 1.0, 1.5) were employed to obtain these results. However, no clear pattern was detected when testing these distinct values, except that a threshold of 0.5 showed higher correlation values. This indicates that fewer cases are more effective in identifying new areas of propagation, but more data and statistical tests are needed to make any further affirmation.

In line with the work of Gilbert et al. [Gilbert et al. 2020], the results confirm that it is possible to determine areas at high risk of infectious diseases using travel probabilities. Furthermore, this approach allowed the development of a visualization tool that permits the final user to explore different perspectives and configurations of the problem. While similar visualization tools exist [EpiRisk 2022, Sadekar et al. 2021], they typically only cover global or other countries' areas. Our tool focuses on Brazilian cities, allowing us to understand the limitations and unique characteristics of the Brazilian data. As

[11] We related each state's confirmed cases to its capital city during the evaluation for verification purposes. This assumption is reasonable because the majority of cases reported are, in fact, related to the capitals and travel flows are directed towards the capital city, with other destinations also including the state capital as part of their itinerary.

a result, we verified that the Brazilian transportation data, particularly the land data, is relatively scarce, which deeply affected the results found in the work. Despite these challenges, our approach offers a lightweight solution for identifying regions at risk.

However, the study's results have limited generalizability due to several constraints. Firstly, there are some gaps in the mobility and validation data. The mobility data used is flawed and likely underestimates the numbers. Sometimes, even data provided by IBGE and transportation agencies, particularly ANTT data, had to be disregarded for not representing reality. Furthermore, the study did not consider car trips, which are highly prevalent in some cities, especially in the countryside. For instance, we believe that adding water transport data can bring better results for the gamma variant since the country's northern region highly uses this type of transport. Regarding the validation data, we noticed that certain records lacked information on the city where the case was reported, only indicating the state. This makes it harder to track disease behavior at a finer level. Secondly, the study assumed traveler's flow to be constant throughout the year. Nevertheless, it is common sense that some flows vary throughout the year, for example, during the holidays. This factor may have affected the validation utilizing data from the gamma variant, as the 2019 transport data used may not match reality during a pandemic situation. Lastly, the evaluation of the computed path probabilities was based on the assumption that COVID-19 was introduced and emerged in the country only from one city. This assumption may be the best alternative for validation; however, at least for the variant alpha, this might not be a reality since Brazilian borders were not closed immediately after the first reported case. These limitations may impact the findings of the present study.

Hence, further research is needed to establish better mobility and validation data throughout Brazil. The methodology of the present study could be replicated with more accurate information, such as using better mobility flow models, geolocalized data from mobile phones, or other databases with additional types of transportation like cars and ferries. The same applies to the validation data. It would be interesting to validate the proposed approach with data from other Brazilian epidemic diseases, such as Zika and Chikungunya. By doing so, we may be able to evaluate the proposed approach better and potentially build a more comprehensive infectious diseases database. The use of graphs for modeling is also an important domain of investigation. There are several algorithms, such as Ford-Fulkerson, Centrality Metrics, and trajectory prediction methods, that can be utilized for graph analysis. These algorithms may provide more effective solutions to the problem at hand. In addiction, the question of how we can integrate disease behavior, for example, with knowledge graphs to make better suggestions and predictions remains to be answered. Researching different visualization types to communicate information to end-users is also necessary. Epiflow is in its initial stages, and there is ample opportunity for improvement in the interface level and model/algorithms used. By doing so, we can create an easy-to-use tool to assist decision-making in tracking disease spread in its early stages and help prevent pandemics.

6. Acknowledgements

This work has been enriched by the invaluable contributions of the following institutions: The ÆSOP (Alert-Early System of Outbreaks with Pandemic Potential) project provided us with the primary idea and guidance for this research, laying the foundation for our

study. CIn-UFPE, SiDi, and Samsung Brazil, who supported the "Data Engineering and Data Science" Residency program, where this study was successfully applied as part of our coursework, culminating in its completion. The collaborative efforts and support from these institutions played a pivotal role in the successful execution and completion of this research project.

References

Bajardi, P., Poletto, C., Ramasco, J. J., Tizzoni, M., Colizza, V., and Vespignani, A. (2011). Human mobility networks, travel restrictions, and the global spread of 2009 h1n1 pandemic. *PloS one*, 6(1):e16591.

Balcan, D., Gonçalves, B., Hu, H., Ramasco, J. J., Colizza, V., and Vespignani, A. (2010). Modeling the spatial spread of infectious diseases: The global epidemic and mobility computational model. *Journal of computational science*, 1(3):132–145.

Berrar, D. (2019). Cross-validation.

Brockmann, D. and Helbing, D. (2013). The hidden geometry of complex, network-driven contagion phenomena. *science*, 342(6164):1337–1342.

Dash (2022). Dash python user guide. "url=https://dash.plotly.com/". Retrieved November 14, 2022.

Dijkstra, E. W. (2022). A note on two problems in connexion with graphs. In *Edsger Wybe Dijkstra: His Life, Work, and Legacy*, pages 287–290.

EpiRisk (2022). Epirisk. "url= https://epirisk.net/". Retrieved November 14, 2022.

Faria, N. R., Mellan, T. A., Whittaker, C., Claro, I. M., Candido, D. d. S., Mishra, S., Crispim, M. A., Sales, F. C., Hawryluk, I., McCrone, J. T., et al. (2021). Genomics and epidemiology of the p. 1 sars-cov-2 lineage in manaus, brazil. *Science*, 372(6544):815–821.

Gilbert, M., Pullano, G., Pinotti, F., Valdano, E., Poletto, C., Boëlle, P.-Y., d'Ortenzio, E., Yazdanpanah, Y., Eholie, S. P., Altmann, M., et al. (2020). Preparedness and vulnerability of african countries against importations of covid-19: a modelling study. *The Lancet*, 395(10227):871–877.

Keeling, M. J. and Rohani, P. (2008). *Modeling Infectious Diseases in Human and Animals*. Princeton University Press.

Ministério da Infraestrutura (2020). Anuário estatístico de transportes 2010 - 2020. "url=https://www.gov.br/infraestrutura/pt-br/assuntos/dados-de-transportes/anuario-estatistico-2". Retrieved November 14, 2022.

Mu, X., Yeh, A. G.-O., and Zhang, X. (2021). The interplay of spatial spread of covid-19 and human mobility in the urban system of china during the chinese new year. *Environment and Planning B: Urban Analytics and City Science*, 48(7):1955–1971.

Nakamura, H. and Managi, S. (2020). Airport risk of importation and exportation of the covid-19 pandemic. *Transport policy*, 96:40–47.

Neo4j (2022). Neo4j. "url=https://neo4j.com/". Retrieved November 14, 2022.

Neo4j GDS (2022). The neo4j graph data science library manual v2.2. "url=https://neo4j.com/docs/graph-data-science/current/". Retrieved November 14, 2022.

Peixoto, P. S., Marcondes, D., Peixoto, C., and Oliva, S. M. (2020). Modeling future spread of infections via mobile geolocation data and population dynamics. an application to covid-19 in brazil. *PloS one*, 15(7):e0235732.

PostgreSQL (2022). Postgresql. "url=https://www.postgresql.org/". Retrieved November 14, 2022.

Reback, J., McKinney, W., Van Den Bossche, J., Augspurger, T., Cloud, P., Klein, A., Hawkins, S., Roeschke, M., Tratner, J., She, C., et al. (2020). pandas-dev/pandas: Pandas 1.0. 5. *Zenodo*.

Sadekar, O., Budamagunta, M., Sreejith, G., Jain, S., and Santhanam, M. (2021). An infectious diseases hazard map for india based on mobility and transportation networks. *arXiv preprint arXiv:2105.15123*.

Statista (2022). Number of flights performed by the global airline industry from 2004 to 2022. "url=https://www.statista.com/statistics/564769/airline-industry-number-of-flights/". Retrieved November 14, 2022.

Wei, Y., Wang, J., Song, W., Xiu, C., Ma, L., and Pei, T. (2021). Spread of covid-19 in china: analysis from a city-based epidemic and mobility model. *Cities*, 110:103010.

WHO (2022a). Who coronavirus (covid-19) dashboard. "url=https://covid19.who.int/". Retrieved November 14, 2022.

WHO (2022b). Who releases 10-year strategy for genomic surveillance of pathogens. "url=https://www.who.int/news/item/30-03-2022-who-releases-10-year-strategy-for-genomic-surveillance-of-pathogens". Retrieved November 14, 2022.