Journal of Information and Data Management, 2025, 16:1, doi: 10.5753/jidm.2025.4291

Applying Graph Databases and Human Mobility Data to Track Infectious Disease Spread in Brazil

Mariama C. S. de Oliveira [**Åbo Akademi** | ***mariama.serafimdeoliveira@abo.fi***]
Lucas Henrique G. de Sales [**Universidade Federal Rural de Pernambuco** | ***lucas.gonzaga@ufrpe.br***]
Andrêza Leite de Alencar [**Universidade Federal Rural de Pernambuco** | ***andreza.leite@ufrpe.br***]
Natalia T. S. de Oliveira [**Universidade Federal de Pernambuco** | ***ntso@cin.ufpe.br***]
Antônio Ricardo K. Cunha [**Fundação Oswaldo Cruz** | ***ricardo.khouri@fiocruz.br***]
Pablo Ivan P. Ramos [**Fundação Oswaldo Cruz** | ***pablo.ramos@fiocruz.br***]

Department of Computing, Universidade Federal de Pernambuco, Rua Dom Manuel de Medeiros, s/n, Dois Irmãos, Recife, PE, 52171-900, Brazil.

Received: 21 March 2024 • **Published:** 22 August 2025

Abstract This study proposes a method for identifying potential routes of disease spread in Brazil, using mobility and graph databases based on open data on cities and transport. The approach incorporates and compares data from different Brazilian monitoring agencies on road, waterway, and air connections, to enhance the understanding of inter-municipal public transport networks. Additionally, the study adapts the Dijkstra algorithm in Neo4j's Data Science module and creates a tool called Epiflow, which helps visual exploration. The proposed approach was validated using COVID-19 data of two variants (alpha and gamma). The results revealed a robust correlation with the alpha variant, whereas inconclusive findings emerged from the gamma variant data. This underscores the hypothesis that the available mobility data in the northern region, which is a rather remote area of the country, could not account for the occurrences of this variant.

Keywords: Graph Database, Human Mobility, Open Data, Infectious Disease, Neo4j

1 Introduction

According to the 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 according to Statista [2022]. In Brazil, about 1.39 million interstate bus trips were performed in 2019 according to 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 article is an extension of the conference paper Oliveira *et al*. [2023] published in the Proceedings of the Brazilian Symposium on Databases 2023. This extended version provides more details regarding the previous study and adds news datasets and experiments aimed at exploring an approach using city and transport data arranged in a graph structure to examine how infectious diseases spread in Brazil. The study also 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 article is organized into six sections. Section 1, Introduction, defines the scope of the study and its significance, along with its objectives. Section 2, Literature Review, presents the prevalent techniques for tracking disease spread. Section 3, Methodology, describes the data, computation techniques used, and the developed application. Section 4, Evaluation and Results, covers the validation process and its outcomes. Section 5, Discussion, discusses the article's discoveries and limitations. Lastly, Section 6, Conclusion, summarizes the study's main findings and proposes 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 organization, COVID-19 has brought to light the implications of such actions by showing the importance of tackling epidemic risk at early stages. In this context, an initiative named ÆSOP[1] (Alert-Early System of

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

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 major challenges the initiative faces is deciding where to collect samples to look for pathogens.

Given this scenario, the current study introduces an approach designed to explore and comprehend the spread dynamics of infectious diseases by leveraging human mobility data and city connections within a graph solution. This endeavor aims to facilitate the formulation of sampling strategies and other public policies geared towards mitigating disease outbreaks at early stages.

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 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*. [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 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, 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*. [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 uses a computationally non-costly approach like the one introduced by Gilbert *et al*. [2020]. Similarly to their work, we use traveling probabilities to track disease-spreading spatial behavior. This allows us to create a tool that provides quick feedback to users, helping them explore different aspects of the problem. Unlike previous studies, which majorly concentrate on air transportation at a global scale, we consider land, air, and water transportation at a Brazilian level. As a result, our analysis should provide a more accurate representation of the spread of disease 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.

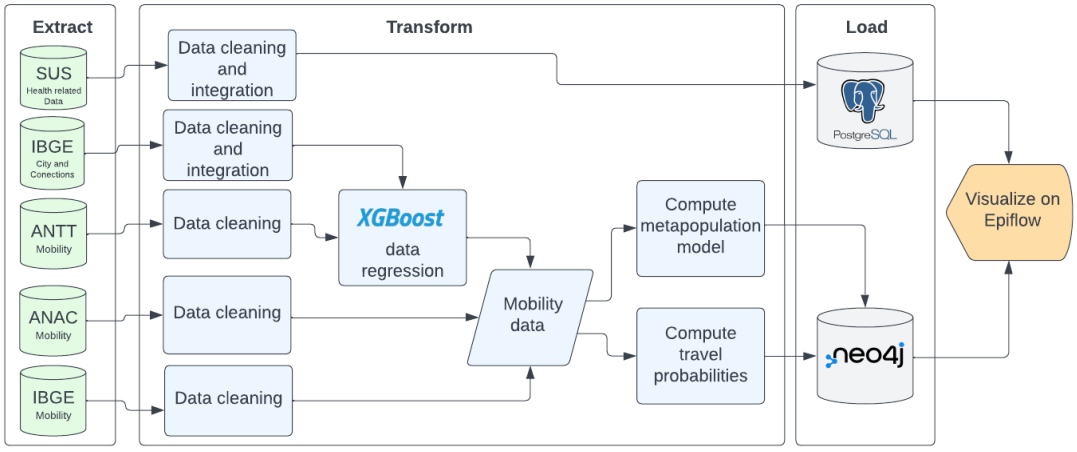


Figure 1. Flow diagram of the ETL methodology adopted.

Table 1. Transport datasets information.

	Transport modalities	Data Sources	No. of connections	No. of municipalities	Transformations
Dataset 1	Road and air	ANTT, ANAC	123,978	4,994	Data Regression
Dataset 2	Road, water and air	IBGE, ANAC	131,646	5,387	Estimation of passengers by vehicle

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. Consequently, it underwent a transformation phase, largely utilizing the Python library Pandas Reback *et al*. [2020]. After the transformations were performed, the data could be utilized in this study. The data characteristics and the transformations performed will be presented as follows. For more details about the data, please consult the project data dictionary[6].

3.1.1 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.

3.1.2 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); 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.

3.1.3 Transport data

Transportation data encompasses information regarding the number of passengers utilizing various public transportation systems between two cities, including road transport (e.g., buses, vans), water transport (e.g., boats), and air transport (e.g., airplanes). To achieve this, we compiled two datasets using different transformations and sources. We opted for two datasets due to distinct limitations inherent to each. By comparing these datasets, our aim is to gain a more comprehensive understanding of the transport networks and to determine which dataset aligns more closely with real-world data. This comparison will facilitate a better grasp of which public data can offer superior assistance in both current and future research endeavors. **Table 1** delineates the characteristics of both datasets. The subsequent sections will delve into the formulation of each dataset and elucidate the essential transformations required for our research.

Dataset 1 The first dataset comprises the data collected from two Brazilian transport regulatory agencies, ANAC and ANTT. Both sources provide information regarding the number of airplane and bus passengers in 2019 within the Brazilian territory. 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 covers a significant portion of passengers who travel using public means in

[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`

[7]The hospitals within this network can also be referred as to Sentinel Service.

the country. However, after a data exploration, it was discovered that the number of passengers between some city pairs was reported inadequately. The reason behind this issue is that ANTT does not require bus companies to collect data on trips that occur within the same state, leading to an underestimation of these records. To address this problem, a machine learning regression model was used to estimate these values. The regressor chosen was the XGBoost regressor[8]. This resulted in the estimation of all bus trips within the same state, as well as some interstate journeys. Note that this issue did not affect airplane data.

Dataset 2 The second dataset comprises data sourced from two Brazilian agencies, IBGE and ANAC. In contrast to the first dataset, this one substitutes the data from ANTT with that of IBGE to ascertain about land passenger flow between cities. The inclusion of IBGE data provides an added advantage as it incorporates information on both road and water transportation, thereby offering insights into the movement of people in regions where water transportation is significant, such as cities located in the northern part of the country.

The data from IBGE was gathered in 2016 through surveys conducted at bus and waterway terminals, as well as at other intercity transport points. This provided a comprehensive view of the intercity public transport network. However, the dataset had one limitation - it did not directly provide the number of passengers that traveled through a connection in a year. Instead, it provided the average number of vehicles (normalized by bus) that traveled through the connection in a week. Therefore, it was necessary to convert the vehicle counts to the number of passengers. This was achieved by multiplying the number of people transported in a vehicle per trip (22) by the average number of weekly trips and the number of weeks in the year[9]. Once again, this type of transformation did not affect the air passenger data since it was already in its correct format.

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 utilized to quickly load general information, such as state, city, and health region names, into the application. At the same time, Neo4j was responsible for modeling the city network and storing the 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.

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 within Brazil, while `HealthRegion` and `ReferenceHospital` store information regarding the Brazilian health system. The arrows in the schema illustrate the relationships between these entities.

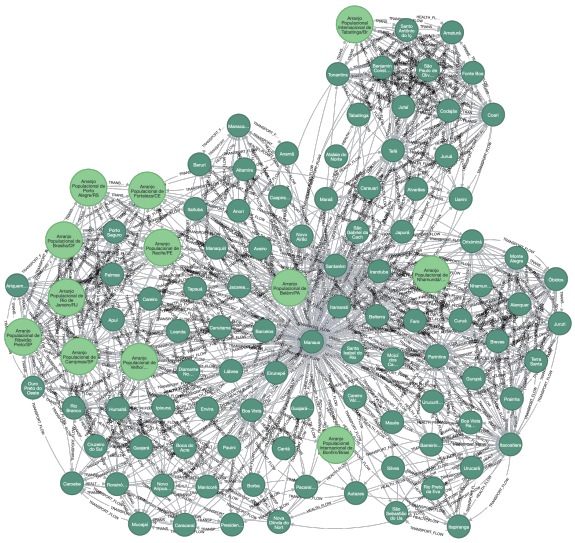
Besides the Brazilian cities, the graph structure also includes other nodes that store information about the studied network, as shown in the graph schema in **Figure 2**; nevertheless, we will focus on the relationship between cities because they are the ones that provide the information needed to identify 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, water and air transportation. This relationship has five attributes: air, bus, water, total (air + bus + water) flow in 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.

3.3 Identifying cities at risk

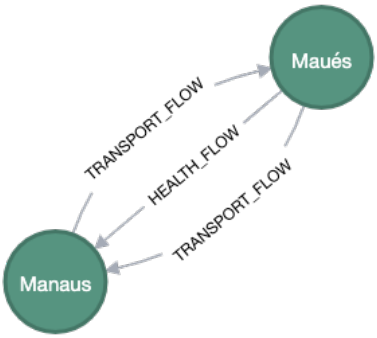
Human mobility plays a vital role in the spread of infectious diseases, as previously discussed. With this premise as a starting point, we define cities at risk as the ones with the highest probability of receiving people from cities where evidence of infectious diseases was found. This approach draws

[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, a total of 19378 observations were used and were divided into a train set (75%) and a test set (25%). A k-fold cross-validation with five folds was used while tuning the model, and ten folds were utilized to obtain the final model's MSE metrics for testing purposes. 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 (10 k-fold) = 0.024; MSE (test set) = 0.409; R^2 (test set) = 0.617.

[9]The number of passengers estimated per vehicle was 22. This value represents the average number of passengers per bus, sourced from the National Department of Transport Infrastructure (DNIT) dataset Orrico Filho *et al*. [2019]. Thus, the calculation of passengers per year is: 52 $\times$ no. of vehicles by connection $\times$ 22. This calculation applies to both road and water transportation, as IBGE has normalized the data to bus terms.



(a)



(b)

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

inspiration from previous projects EpiRisk [2022]; Gilbert *et al*. [2020] due to being a computationally inexpensive method of determining areas susceptible to pathogen propagation. Calculating this probability is straightforward: it entails understanding the volume of people traveling from one city to another, as depicted in Equation 1. Leveraging data obtained from transport agencies and regression analysis, we were able to compute the probability of travel between cities. This probability also serves as an indicator of the risk that a particular city may face if evidence of pathogens is detected in one of its connections.

$$Pr(\text{A to B}) = \frac{\text{No. of passengers from city A to city B}}{\text{Total no. of passengers from city A}} \quad (1)$$

While the computation presented in equation 1 was necessary to determine traveling probabilities based on passenger flow, the flow of patients using health services in another city 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 discussed further in the next section.

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 probabilities of travel, thereby allowing us to identify the path with the highest likelihood of occurrence. This computation is facilitated by Dijkstra's algorithm Dijkstra [2022], initially designed to find the shortest path between two nodes in a weighted graph. For Dijkstra's shortest path search algorithm, three factors are crucial: edge costs, total cost, and the function used. The edge cost represents the weight of each edge, the total cost is the summation of all edge weights along the path (initially 0, increasing as we traverse the graph), and the function employed is typically summation. 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 probability values, and we want to calculate the probability of events happening at the same time).
2. The initial cost, instead of being 0, was -1 (to prevent multiplication resulting in 0 and to 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.

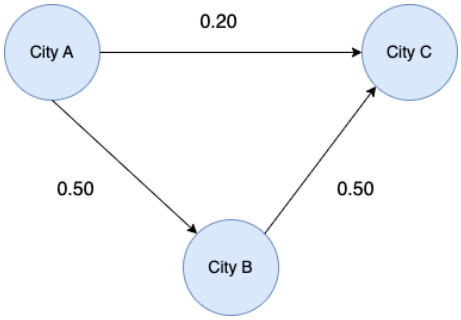


Figure 4. The directed graph shows three connected cities with their traveling probabilities.

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 [10] to add the mentioned changes.

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

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)**).

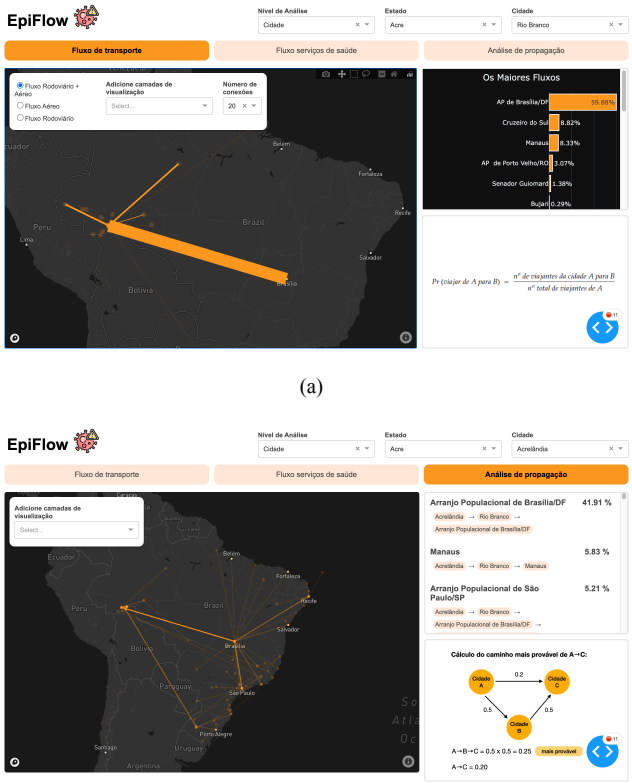


Figure 5. Epiflow screens: (a) Visualization of the transport flow. (b) Visualization of spreading routes to specific cities.

In the screen shown in **Figure 5(a)**, the Epiflow app indicates which cities are at higher risk if some disease starts spreading from Rio Branco in the state of Acre. Brasília (55.88%), Cruzeiro do Sul (8.82%), and Manaus (8.33%) emerge as the top three at-risk cities as we can see in the barplot. Despite not sharing borders with Rio Branco, these cities receive a significant flow of people from it, according to the data.

The subsequent screen (**Figure 55(b)**) of the app displays the main paths that a disease would likely take if it started spreading from Acrelândia. For instance, the route Acrelândia → Rio Branco → Brasília was determined to be the most probable, with a 41.91% likelihood. This makes sense because there is a notable movement of people from Acrelândia to Rio Branco and from Rio Branco to Brasília in the utilized data.

As demonstrated, after selecting the state and city, the app can present the cities at higher risk and the most probable transmission routes.

4 Evaluation and Results

After implementation, we evaluated our approach with data series from two COVID-19 variants highly disseminated 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 of this variant. That is an assumption that we will hold during this evaluation. Regarding the gamma variant, we will consider its epicenter as the city of Manaus, as this variant originated in this region, according to genetic surveillance Faria *et al.* [2021].

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[11] 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 not sufficient 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 2**.

According to **Table 2**, all the computed values for the variant 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. These remarks apply to both flow datasets.

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, the study employed Spearman's rank correlation to determine the correlation between the ranking of capital cities[12] (explained in sub-

[11] The calculation of the moving average used a window size of 7. This method aims to reduce anomalies in the reported case numbers over time. Exploratory data analysis performed in the current study revealed that there is a concentration of reported cases on certain days of the week due to health agencies' procedures, which could jeopardize the evaluation of the results.

[12] 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 at the beginning of the pandemic are, in

Table 2. 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
Dataset 1	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
Dataset 2	Alpha	r(24) = -.86, p = .000	r(24) = -.82, p = .000	r(24) = -.84, p = .000
	Gamma	r(11) = -.40, p = .099	r(11) = -.27, p = .281	r(11) = -.27, p = .284

Table 3. 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
Dataset 1	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
Dataset 2	Alpha	r(24) = -.83, p = .000	r(24) = -.84, p = .000	r(24) = -.77, p = .000
	Gamma	r(24) = -.28, p = .165	r(24) = -.13, p = .523	r(24) = -.09, p = .664

section 4.1) and the likelihood of a disease spreading along a particular route. The results are found in **Table 3**.

In **Table 3**, for both datasets, there is a strong correlation between the likelihood of disease transmission through a particular route and the actual data of the alpha variant. Additionally, for Dataset 1 using a threshold of 0.5 cases, 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 on Dataset 1. All the other values for the variant gamma are inconclusive since they are not statistically significant ($p > .05$).

5 Discussion

Based on the collated results, some interesting conclusions can be drawn regarding identifying cities at risk and disease propagation routes within the Brazilian territory. This section will reflect upon these findings and the research process. Furthermore, the work's implications and limitations will also be discussed.

The evaluation results on both datasets show a clear correlation between human mobility (computation of 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 the lack of statistically significant results. Regarding the propagation route (pathways), there is a moderate to strong correlation between the study's estimated probabilities and the order in which the alpha variants first spread to the Brazilian capitals. Nevertheless, once again, we were not able to conclude in this area while validating with the gamma variant.

Regarding the different moving average thresholds (0.5, 1.0, 1.5) utilized to evaluate the two proposed approaches, we were only able to identify a threshold of 0.5 yielded higher correlation values. This suggests that using fewer cases might be more effective in identifying new propagation areas. However, it is essential to note that further data and additional statistical tests are required to draw any definitive conclusions. These findings underscore the complexity of understanding the dynamics of pathogen spread and refining methodologies for this purpose.

Another relevant aspect of research is the two datasets utilized. In this area, we observed two main findings. The first finding was that the analysis of both datasets yielded comparable outcomes for assessing alpha and gamma variations. Notably, strong correlations were evident for the alpha variant across both datasets, while findings for the gamma variant were inconclusive. These results surprised us, as we had anticipated a significant correlation in Dataset 2 for evaluating the gamma variant, given that this variant emerged in a region highly dependent on water transportation. However, we presume that the scarcity of information regarding road connections in the northern region of Dataset 2 may have contributed to the diminished correlation for the gamma variant despite integrating data from both road and waterway networks. The second finding is that the comparable outcomes across both datasets underscore their similar coverage and portrayal of inter-municipality flow dynamics. This suggests that transformations applied, such as data regression employing XGBoost in Dataset 1, did not significantly alter the data behavior. As a result, these transformations can serve as a valuable resource for rectifying issues stemming from missing or under-reported data in comparable datasets.

Despite some unexpected results, the findings are in line with the work of Gilbert *et al*. [2020]. The results confirm that it is possible to determine areas at high risk of infectious diseases using travel probabilities. Furthermore, the current study 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 dynamics. As a result, we verified that the Brazilian transportation data, particularly the land and water data, is relatively scarce, which deeply affected the results found in the study. Despite these challenges, our approach offers a lightweight solution for identifying regions at risk.

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.

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 locations, especially in the countryside. For instance, we believe that adding water transport data with more precise values for the number of people traveling by water can bring better results for the gamma variant since the country's northern region highly uses this type of transport. Furthermore, while the study addresses disease transmission due to transportation, it does not examine the dynamics of disease transmission within confined spaces like buses and airplanes, where individuals share close quarters for extended periods and can potentially spread diseases. Although this is an important aspect to consider, the primary focus is on how transportation can facilitate the spread of diseases by facilitating the movement of people from one city to another.

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, particularly at the municipality level. Secondly, the study assumed traveler's flow to be constant throughout the year. Nevertheless, it is common sense that some flows are seasonal, varying throughout the year, for example, during the holidays. This factor may have affected the validation with the gamma variant, as the 2019 transport data used may not match reality during a pandemic. 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.

6 Conclusion

The findings of this study emphasize the importance of using human mobility to identify high-risk areas for infectious diseases and track their spread in Brazil. By analyzing Spearman's correlation, we discovered a direct link between the system's travel probabilities and the spread of the alpha variant across different Brazilian states. However, due to data limitations, the correlation for the gamma variant was inconclusive.

Moreover, we found that using multiple datasets is crucial to overcome issues like missing data and underrepresented information. However, while evaluating the gamma variable, we concluded that there is still room for improvement in obtaining data on human mobility in some parts of Brazil, particularly less density areas like the north of Brazil.

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. 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 addition, the question of how we can integrate disease behavior with knowledge graphs to make better suggestions and predictions remains to be answered. Furthermore, researching different visualization types to communicate results information to end-users is also necessary.

The current study and developed application are in their initial stages, and there is ample opportunity for improvement in the interface level, model and 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.

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 UFRPE collaborative efforts and orientation support played a pivotal role in the successful execution and completion of this research project.

Authors' Contributions

ALA, MCSO, ARKC and PIPR contributed to the conception of this study. MCSO and LHGS performed the experiments. MCSO and NTSO contributed to the app implementation. All authors read and approved the final manuscript.

Availability of data and materials

The application code and datasets generated and analyzed during the current study are available in urlhttps://github.com/mariamaOlive/alerta-pandemia.

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.

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.

Oliveira, M., Alencar, A., Oliveira, N., Sales, L., Cunha, A., and Ramos, P. (2023). Epiflow: a hybrid approach to track infectious disease spread in brazil based on travel data and graph databases. In *Proceedings of the 38th Brazilian Symposium on Databases*, pages 218–230, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/sbbd.2023.231736.

Orrico Filho, R. D., DNIT, and COPPE/UFRJ (2019). Dados de origem e destino das pesquisas nacionais de tráfego de 2016 e 2017. Nota Técnica 04/2019/DE, Departamento Nacional de Infraestrutura de Transportes (DNIT), Rio de Janeiro. Termo de Execução Descentralizada - TED 964/2014 — DPP.

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.