

Oxford Internet Institute, University of Oxford

Assignment Cover Sheet

Candidate Number	1053722
Assignment	SDS MSc Thesis
Term	Trinity Term 2021
Title/Question	Slowing the Spread: Assessing Early COVID-19 Intervention Effectiveness in São Paulo using Socio-Demographic and Mobility Data
Word Count	14997

By placing a tick in this box ☑ I hereby certify as follows:

(a) This thesis or coursework is entirely my own work, except where acknowledgments of other sources are given. I also confirm that this coursework has not been submitted, wholly or substantially, to another examination at this or any other University or educational institution;

(b) I have read and understood the Education Committee's information and guidance on academic good practice and plagiarism at `https://www.ox.ac.uk/students/academic/guidance/skills?wssl=1`.

(c) I agree that my work may be checked for plagiarism using Turnitin software and have read the Notice to Candidates which can be seen at: `http://www.admin.ox.ac.uk/proctors/turnitin2w.shtml`, and that I agree to my work being screened and used as explained in that Notice;

(d) I have clearly indicated (with appropriate references) the presence of all material I have paraphrased, quoted or used from other sources, including any diagrams, charts, tables or graphs.

(e) I have acknowledged appropriately any assistance I have received in addition to that provided by my [tutor/supervisor/adviser].

(f) I have not sought assistance from a professional agency;

(g) I understand that any false claims for this work will be reported to the Proctors and may be penalized in accordance with the University regulations.

Please remember:

- To attach a second relevant cover sheet if you have a disability such as dyslexia or dyspraxia. These are available from the Higher Degrees Office, but the Disability Advisory Service will be able to guide you.



University of Oxford

Slowing the Spread: Assessing Early COVID-19 Intervention Effectiveness in São Paulo using Socio-Demographic and Mobility Data

1053722

Thesis submitted in partial fulfilment of the requirement for the degree of MSc in Social Data Science at the Oxford Internet Institute at the University of Oxford

Trinity Term 2021

14997

Acknowledgements

Thank you to Chico for being incredibly kind and supportive throughout this entire process. I am so grateful to have spent my year working with you and – thesis aside – you have taught me so much about being curious, open minded, and embracing meaningful challenges. I would also like to thank Pedro and Rafa for their insightful feedback and discussions, crucial to creating this project. Finally, I would like to thank my mother, father, and brother for all of their support. Spending months at home with you was not in my original plans, but turned out to be a highlight of the year.

Abstract

Background and Motivation: The COVID-19 pandemic continues to have devastating health, social, and economic impacts around the world. Brazil is second globally in COVID-19 deaths, with spiralling growth after its first case in the Metropolitan Region of São Paulo (MRSP), the largest urban agglomeration in South America. Lockdown interventions to reduce mobility are a leading tool for infection mitigation, and while effective on aggregate, their adherence is impacted by socio-demographic factors. A forward-looking model that evaluates the ability of lockdowns to slow disease spread through a socio-demographic lens can inform more equitable policy-making.

Methods: Cellphone mobility data integrated with traditional commuting survey, census, and COVID-19 data from the MRSP are used to develop a novel effective distance-based model to measure the ability of lockdowns to slow disease spread. This Infection Delay Model (IDM) is applied to simulated epidemics across the MRSP. The first analysis clusters *regions-at-risk* that benefit similarly from lockdowns. The second analysis identifies *outbreak regions* where interventions lead to high and low slowdowns for the rest of the MRSP. Socio-spatial context provides meaningful and actionable interpretations of lockdown effectiveness.

Results: For regions-at-risk, the clustering analysis reveals that socio-demographically vulnerable regions with lower centrality have greater delays to their first case by early lockdowns. The second analysis, observing each region as an initial source of infection, reveals that outbreaks beginning in socio-demographically secure regions with higher centrality are disproportionately slowed by lockdowns to the rest of the MRSP.

Conclusion: A fundamental relationship exists between social inequality and the ability of lockdowns to slow infection spread. First cases in socio-demographically vulnerable regions can be most delayed by early lockdowns, conveying that rapid interventions can reduce health inequities. The outbreak analysis shows that lockdowns slow the spread more for outbreaks beginning in more secure regions. Increasing capacity of vulnerable regions to isolate can strengthen the MRSP's resilience to future outbreaks.

Keywords: Socio-Demographic Inequality; Human Mobility Data; Epidemiological Modelling; Spatial Analysis; Unsupervised Learning; Public Policy

1 Introduction

1.1 Background

COVID-19 is a deadly respiratory disease that first appeared in Wuhan, China in December 2019. By January 30, 2020 it was declared a Public Health Emergency by the World Health Organization (WHO), and by 11 March, 2020, its uncontainable spread led to its declaration as a pandemic (Lai et al., 2020). According to the Johns Hopkins COVID-19 Data Repository, as of August 2021, COVID-19 has killed over 4 million individuals with 195 million confirmed infections worldwide (Dong, Du, & Gardner, 2020). To mitigate its spread, non-pharmaceutical interventions (NPIs) such as mask mandates and lockdowns have become vital tools of governments and local authorities around the world (Haug et al., 2020). More virulent variants and fears regarding future epidemics have focused scholarly attention on maximising the effectiveness of NPIs and containing the spread of deadly infections (Li, Sabrina, Pereira, Prete, Zarebski, & Emanuel, 2020; Lee, Qian, & Schwanen, 2020; Mascola, Graham, & Fauci, 2021). In addition to the efficacy of these measures, the social equity of NPIs has become a critical question for researchers striving for equal protection among all segments of society.

The first COVID-19 case in South America appeared in Brazil on February 25, 2020. Brazil rapidly became an epicenter and has since been devastated by the disease (Ribeiro, Sunahara, Sutton, Perc, & Hanley, 2020; F. S. H. de Souza, Hojo-Souza, Batista, da Silva, & Guidoni, 2021). As of July 31 2021, it ranked second for overall deaths behind the United States, with emerging variants circulating throughout the country (Imai et al., 2021). Brazil is a major international transport hub with a population of 211 million (Burns, Martins, Schneider, Momsen, & James, 2021). Combined with its rapid case growth and variant propagation, Dr. Miguel Nicolelis, a Brazilian professor of medicine explained to the BBC on 7 April 2021:

"If Brazil is not under control, then the planet is not going to be safe, because we are brewing new variants every week... and they are going to cross borders" ("Covid: Brazil has more than 4,000 deaths in 24 hours for first time", 2021).

Brazil's first case occurred in the Metropolitan Region of São Paulo (MRSP), the largest urban agglomeration in South America (Neiva et al., 2020; Coy, Sandholz, Töpfer, & Zirkl, 2018). The geography of the MRSP within São Paulo state is presented in Figure 1. The MRSP has a population of over 22 million spread across 39 municipalities (Candido et al., 2020). The MRSP was home to half of the

initial international COVID-19 arrivals to Brazil, putting it at the highest risk of outbreaks (Candido et al., 2020). Given its unique position as an area of early disease introduction and high intrastate transmission, COVID-19 studies in the MRSP can help to prepare for future outbreaks (Candido et al., 2020).

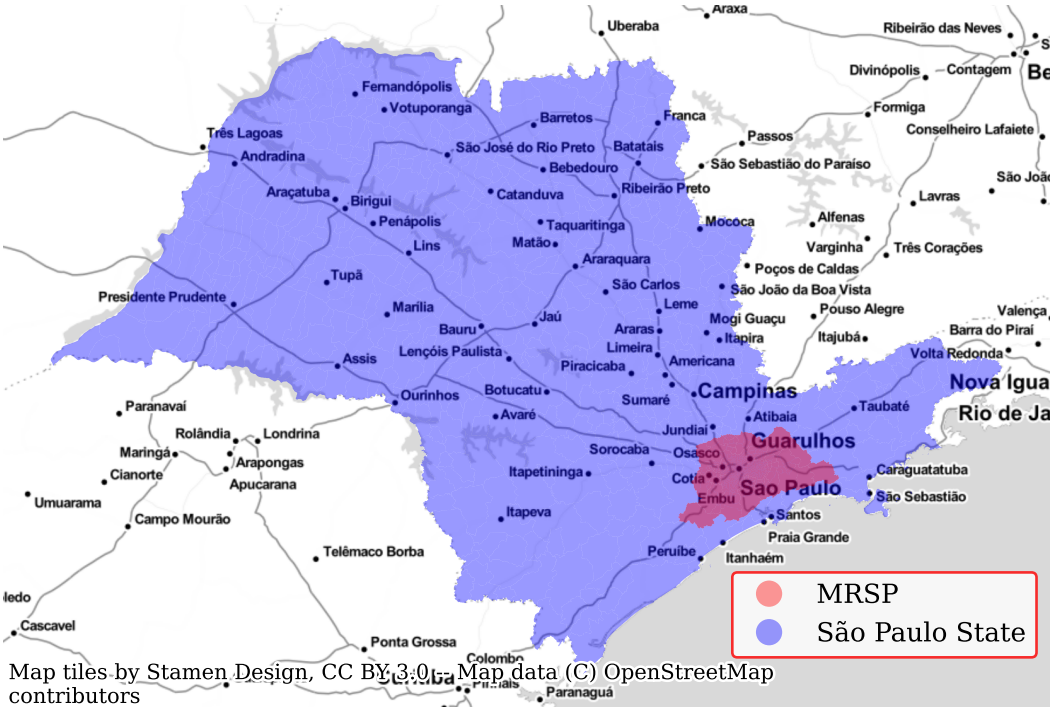


Figure 1: Geographic representation of regions within the MRSP compared to the entire state of São Paulo

The Brazilian Public Health System is the largest universal health system in the world, supporting 80% of the population, and is extremely vulnerable to breakdown if large proportions of the population rapidly become infected (de Andrade, Pereira, Martins, Lima, & Portela, 2020; Neiva et al., 2020). Early interventions, such as mobility restrictions, are invaluable tools to contain the transmission of a novel disease and slow down its initial spread (Neiva et al., 2020). Despite implementation of strict regional lockdown measures on March 16, 2020, within 19 days of the first initial case, outbreaks in the MRSP persisted (Savaris, Pumi, Dalzochio, & Kunst, 2021). Figure 2 displays the mean social isolation in relation to the March 16 lockdown.

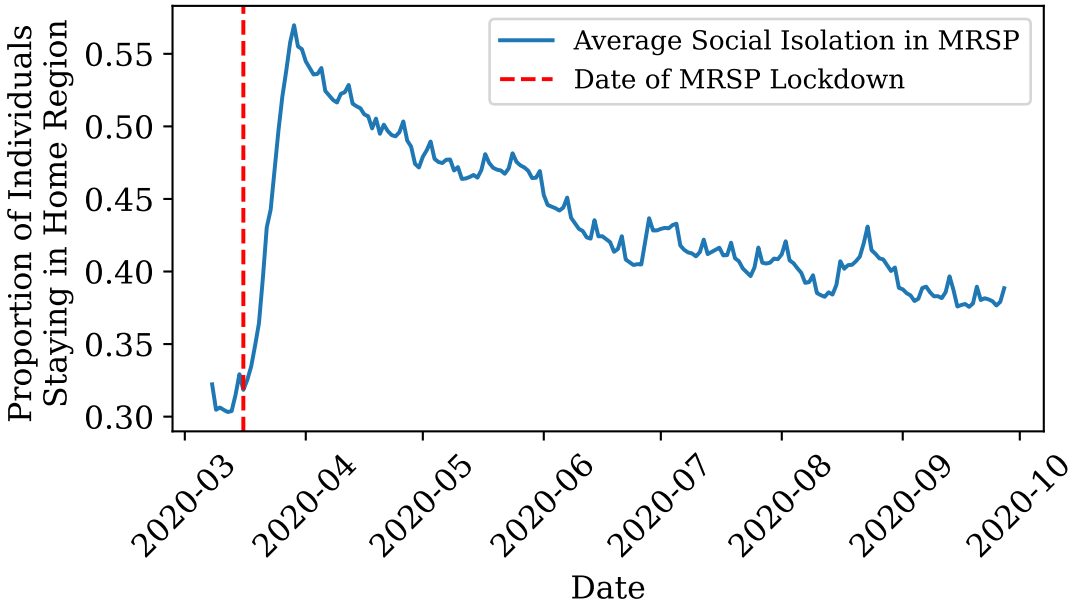


Figure 2: Mean proportion of individuals in the MRSP that stayed in 740m^2 region around their home on a given day, used as a proxy for social isolation (7-day rolling average). March 16, 2020 lockdown indicated with dotted vertical red line.

The humanitarian disaster of COVID-19 warrants urgent research into the effectiveness of interventions to slow the spread of disease. Socio-demographic and health inequalities in Brazil have a substantial impact on the harms caused by COVID-19 and novel variants (Malta, Murray, da Silva, & Strathdee, 2020). Inequality and poverty are driving factors for high transmission and mortality rates in Brazil (Henry, 2020; Chakraborty, Bhattacharya, Sharma, Lee, & Agoramoorthy, 2021). COVID-19 risks amplified by inequalities are partially explained by reduced intervention adherence, specifically for lockdowns (Lee et al., 2020). A growing body of research on the impacts of such interventions focuses not only on aggregate outcomes in Brazil, but on the minimisation of health gaps along socio-demographic dimensions (Landmann-Szwarcwald & Macinko, 2016). Current policy-makers lack detailed information on the effectiveness of early lockdowns to slow the spread of disease in Brazil, and how this intersects with socio-demographic inequalities.

This paper assesses the effectiveness of early lockdowns in the MRSP and their intersection with regional socio-demographic inequalities, shedding light on policy impacts across groups of varying vulnerability. To assess intervention effectiveness, this study develops a novel effective distance-based model, referred to as the Infection Delay Model (IDM), which captures the difference between disease arrival times with and without a lockdown – and how this changes over the course of an

outbreak. The infection delay of an intervention is the primary metric used to assess effectiveness. Instead of determining intervention arrival times with uniform mobility reductions, a new application of cellphone mobility data for effective distance research is developed for the IDM, integrating granular real-world mobility reductions after the March 16 lockdown. The paper simulates 2737 epidemics beginning in all available regions of the MRSP by using commuting-based compartmental epidemiological models. Outbreak simulations across the MRSP ensure the relevance of results given the possibility that future outbreak scenarios may look different from that observed during COVID-19. The approaches of this study, requiring the integration of cellphone mobility, commuting, COVID-19, and census data, create forward-looking insights of lockdown effectiveness to help slow future disease outbreaks.

The first analysis clusters regions *at risk of infection* based on their infection delays (time gained before a first case) after lockdown – taking into account 2737 simulated outbreak scenarios. These clusters are then analysed based on their socio-demographic and commuting network centrality characteristics. This can be summarised by the following research question:

- Research Question 1: When *at risk of infection*, how do regions whose first case is delayed the most and least by a lockdown differ on socio-demographic characteristics?

The second analysis evaluates intervention effectiveness from the perspective of outlying outbreak locations, in the form of the following question:

- Research Question 2 : What are the socio-demographic characteristics of the *initial outbreak regions* where lockdowns lead to disproportionately high and low infection delays for the rest of the MRSP regions?

The analysis has two primary hypotheses regarding the social implications of interventions, expanded upon in the literature review section. The first hypothesis concerns regions *at risk* that benefit the most from early interventions:

- H_1: The first case of infection in less central and more socio-demographically vulnerable regions will be most delayed by an early lockdown.

The second hypothesis concerns outlying regions *of outbreak* in which disease spread across the MRSP can be delayed to the greatest extent:

- H_2: Outbreaks beginning in more socio-demographically secure regions will be slowed more by an early lockdown.

The unique approaches taken in this paper to discern lockdown effectiveness are computationally efficient, replicable, and interpretable. The results provide socially relevant insights into the effectiveness of lockdown measures.

This paper begins with a literature review that centres its analyses within broader mobility, epidemiology, and inequality-related research in Brazil and elsewhere. The methods are then presented, beginning with a description of the data sets and pre-processing steps. This is followed by an explanation of how infection delays are measured using the IDM, how outbreak simulations are generated, and how effective distance metrics are validated for use in the arrival time calculations. The Methods section then discusses how regions in the MRSP are clustered and characterised based on lockdown effectiveness, and how outlying regions of outbreak that lead to above and below average lockdown effectiveness are identified and analysed. The Results section starts by validating effective distance metrics to predict arrival times on the MRSP commuting network. Subsequently, three clusters of regions-at-risk are identified that respond similarly to mobility-based interventions in the MRSP – taking into consideration all possible outbreak scenarios. These clusters are evaluated by the benefit conferred by early lockdown to delay infection onset and their socio-demographic and centrality features. In addition, the Results report on the regions of outbreak in which a lockdown is least and most effective in slowing spread across the MRSP. These results are interpreted in relation to their socio-demographic and commuting network centrality metrics. The Discussion section explores the intersection of inequalities with lockdown effectiveness in the MRSP, what this means for future outbreak preparation, and the limitations, future directions, and original contributions of this study.

1.2 Literature Review

Effective Distance

The arrival times of COVID-19 under baseline and intervention mobility scenarios are calculated using effective distance, a network-based concept. The effective distance between two regions is represented by the flow of individuals travelling between them, rather than a conventional geographic distance (Coelho et al., 2020; Brockmann & Helbing, 2014). To illustrate, if two neighbouring regions are close together but have minimal commuting flow between them, they are effectively 'far'. Similarly, if they are on opposite ends of the MRSP but have high rates of population flow, they are effectively 'close'. Used in numerous epidemiological studies, this measure is highly predictive of infection arrival times, providing insights into dis-

ease transmission dynamics (Coelho et al., 2020; Gautreau, Barrat, & Barthélemy, 2008). Brockmann and Helbing (2014) found effective distance metrics on global airline transportation networks to be highly associated with arrival times of H1N1 and SARS, far exceeding predictions based on geographic distance measures. Their finding was also extended to the Enterohemorrhagic E. coli outbreak in Germany using an effective distance analysis on their food transport network (Brockmann & Helbing, 2014). This generalisation of 'distance' is entirely dependent on the mobility network of a population, and determines arrival rankings which can be translated to arrival *times*, explained in the Methods section. For the purposes of this analysis, the MRSP commuting network is highly suitable for an effective distance analysis.

Daily Commuting and Disease Transmission

Epidemiological studies highlight the importance of daily commuting flow as a vector of disease transmission (Wen, Lin, & Fang, 2012; Francetic & Munford, 2021). Commuting introduces numerous opportunities for the spread of infection, whether from environmental factors, co-workers, or exposures en route (Francetic & Munford, 2021). Francetic and Munford (2021) found that regions with more individuals commuting in and out experience higher COVID-19 mortality rates. Their spatial regression model for COVID-19 in England showed that a 10% increase in the proportion of individuals from one region commuting to other regions would increase their COVID-19 deaths by 2.75 per 100,000. Another key finding, further justifying an investigation of early intervention effectiveness, was that the association between commuting flow and mortality is strongest at the outset of the pandemic (Francetic & Munford, 2021).

The importance of intrastate commuting for the diffusion of COVID-19 was further explored in Minnesota by Levin, Shang, and Stern (2021). Minnesota has minimal interstate commuting and low international airline travel – allowing for a less contaminated assessment of the effects of daily commuting on disease propagation (Levin et al., 2021). Using a network-based compartmental epidemiological model, they found that the real peak of 6,982 COVID-19 cases would have risen to 11,028 without reductions in daily commuting (Levin et al., 2021). Emphasis was placed on short-term travel, such as home-work commuting, where reductions led to substantial decreases in the number of cases and its geographic spread (Levin et al., 2021). These findings justify a commuter-oriented analysis of disease arrivals in the MRSP, a highly dense and interconnected metropolitan area (Coelho et al., 2020).

Inequalities in Brazil Contribute to Disease Burden

Brazil, the ninth largest economy in the world, is also one of the world's most unequal, with 20% of its population living in poverty (Malta et al., 2020). The systemic inequalities present in Brazil have been further exposed by the ongoing pandemic, spreading rapidly in poor areas lacking adequate health care (Malta et al., 2020). To illustrate the stark inequality and resource gap in Brazil, roughly 12 million individuals live in *favelas*, without sanitation or functioning water networks (Malta et al., 2020). These disparities are longstanding, and were apparent during the dengue, Zika, and H1N1 outbreaks (Malta et al., 2020). For future epidemics, where similar health burden gaps are likely to exist, understanding regional patterns of intervention effectiveness can provide information on how to better protect those most vulnerable in a population.

To understand patterns of intervention effectiveness across the MRSP, the inclusion of socio-demographic data is essential. More than population age structure and health risk prevalence, Rocha et al. (2021) found that socio-economic vulnerability was the most important factor in COVID-19's initial spread in Brazil. Delaying disease onset with early interventions can allow health systems to increase hospital capacity, intensive care units, and establish rapid testing sites (Rocha et al., 2021). Further, economic interventions by governments can increase the ability of vulnerable populations to adhere to mobility-based restrictions (Wright, Sonin, Driscoll, & Wilson, 2020). Despite their efficacy, the ability of states to implement them varies.

In Brazil, inadequately funded health systems, and health and socio-demographic inequalities have been found to impact the speed and comprehensiveness of health and economic measures (Rocha et al., 2021). More vulnerable regions were found to have fewer hospital resources, such as ICU beds and physicians per capita (Rocha et al., 2021). This was especially important in the early phase of the pandemic, as vulnerable populations had higher acceleration of deaths per 100,000 individuals prior to June 2020 (Rocha et al., 2021). The sharp inequalities of early outbreak outcomes in Brazil motivate the study of how lockdown timing can mitigate adverse impacts on more vulnerable members of the population. This analysis considers vulnerable populations to be those of lower income and functioning water networks, and greater black and brown populations proportions, based on prior literature's emphasis on socio-demographic intersections with health outcomes in Brazil (Li, Sabrina, Pereira et al., 2020; Hone et al., 2021; Coelho et al., 2020).

The intersection between daily commuting and Brazil's inequalities provides the basis for H_1: the first case of infection in less central and more socio-demographically vulnerable regions will be most delayed by an early lockdown. According to Vasconcellos

(2017), more socio-demographic vulnerable regions in the MRSP have lower commuting rates, due to far more limited transport options. These regions also have fewer job opportunities, and thus face lower demand from outside commuters (Vasconcellos, 2017). Together, these two forces lead to lower centrality for more socio-demographic vulnerable regions on a commuting network, and hence a more delayed disease arrival, on average, if mobility restrictions are implemented early.

Cellphone Mobility Data to Study COVID-19

Cellphone mobility data is an essential tool to study human population movements during the ongoing pandemic (Li, Tian, Lai, & Zhang, 2011; Peixoto, Marcondes, Peixoto, & Oliva, 2020; Chang et al., 2020; Wright et al., 2020). The step change in spatio-temporal granularity relative to traditional census and survey data grants it the status of Big Data by Schroeder's (2014) definition. As a disease spreading primarily through respiratory droplets in proximity to an infected individual, understanding contact patterns is essential (Lv et al., 2020; The Lancet Respiratory Medicine, 2020). The fine spatio-temporal resolution of cellphone mobility data is especially important given the alterations in human mobility patterns caused by COVID-19 (Chang et al., 2020). NPIs have been a major containment strategy of governments around the world to limit the spread of COVID-19. Often, such interventions entail restrictions on mass gatherings, school closings, an emphasis on remote working over commuting, and bans on voluntary trips (Schlosser et al., 2021).

To model the spread of disease and provide strategic information to governments on the efficacy and equity of NPIs, cellphone mobility data has been used as an efficient and timely tool in Brazil (Peixoto et al., 2020). Peixoto et al. (2020) use cellphone mobility data to determine foci of potential infection from travel routes between the states of São Paulo and Rio de Janeiro, and to assess whether the mobility restrictions lead to changing patterns in movement between the states (Peixoto et al., 2020). Aggregated location pairs from March 2019 are used as a pre-pandemic control-group, and compared against movements during the pandemic to assess changing patterns (Peixoto et al., 2020). These positional recordings are used to model the evolution of COVID-19 in pre-pandemic and pandemic mobility scenarios, ranking the regions most at risk and informing local health-care authorities of potential outbreaks.

Integrating Cellphone and Traditional Survey Data Sets

While mobile phone data sets provide up-to-date fluctuations in mobility to capture pandemic-related changes, information on the sample of users represented in the data are often lacking (Tizzoni et al., 2014). While fundamental coverage issues exist in mobile phone data-sets, they can be used in conjunction with traditional census and survey data to answer questions of a social nature (Peixoto et al., 2020). This method of adding depth to mobility data has been considered in numerous COVID-19 studies. Lee et al. (2020) link cellphone mobility data with the UK's 2011 Census to understand the relationship between regional demographic variables and ability to self-isolate. While the analysis of cellphone mobility data indicates that the median displacement across the UK had reduced by 70%, measured using the median radius of gyration, there are strong regional variations in mobility reduction. One of the strong relationships detected is a positive relationship between socio-economic status and ability to self-isolate (Lee et al., 2020). Their timely analysis points to social isolation and safety measures as a 'luxury' that not all UK residents could afford, with important implications on the equity of government responses to COVID-19.

In an analysis of the MRSP, Li, Sabrina, Pereira et al. (2020) attempt to discern the social and racial inequalities in COVID-19 hospitalisations and deaths. Using census data and a Brazilian database on severe acute respiratory illnesses (SARI), the paper finds that black or brown Brazilians had a higher probability of both hospitalisation and death relative to white individuals (Li, Sabrina, Pereira et al., 2020). This inequality can be partially explained by higher comorbidities and lower access to private health care among these disadvantaged groups. Li, Sabrina, Pereira et al. (2020) spatially integrate their census, SARI, and cellphone mobility data in the MRSP to analyse differences in social and racial groups' ability to self-isolate. The cellphone data is structured as a grid of 740m^2 hexagons across the MRSP, each region providing the daily proportion of users who reside in a hexagon and stay on that day. Using these hexagons as proxies for adherence to social isolation guidelines, the authors find that wealthier and whiter neighbourhoods were able to isolate faster and for longer. In contrast, low-income and black and brown individuals, who disproportionately work in precarious, day-to-day jobs without the ability to telework or receive support payments, were less able to limit their mobility (Li, Sabrina, Pereira et al., 2020). The observed hospitalisation and death toll was partially explained by these mobility differences – an insight made possible by using traditional and cellphone mobility data in tandem.

Building on the work of Li, Sabrina, Pereira et al. (2020), this thesis develops H_2: Outbreaks beginning in more socio-demographically secure regions will be

slowed more by an early lockdown. Using cellphone mobility data in the MRSP, the intervention scenarios in this analysis are able to capture regional heterogeneity in social isolation changes. This thesis uses the same data set as Li, Sabrina, Pereira et al. (2020), who find that more affluent regions were able to self-isolate more than vulnerable regions. The analysis employed extends their findings by fusing cellphone mobility, commuting, and census data to examine how these inequalities translate into concrete differences in mitigation. This analysis anticipates greater slowdowns across the MRSP in scenarios where affluent regions are the outbreak locations, due to their greater ability to reduce mobility.

Assumptions of COVID-19 Initial Conditions

A primary contribution of this paper is a forward-looking, generalised model of disease risk, addressing a gap in the literature on COVID-19. Existing literature on modelling COVID-19 assumes the same initial conditions provided by the outbreak. For example, Chang et al. (2020) use a metapopulation model and the mobility patterns observed during the pandemic to model the spread of COVID-19 across metropolitan areas in the United States. The model identifies intervention effectiveness, higher infection rates among disadvantaged individuals, and key places of interest for super-spreader events (Chang et al., 2020). While useful for understanding the events that transpired during COVID-19, their findings are highly context specific, and do not necessarily hold for novel variants or future pandemics where the origin location and spreading patterns may vary.

Brockmann and Helbing (2014) conduct an effective distance analysis on a global airline network's transportation ties to provide simplified models of disease spread. They find that, given a known outbreak location of a disease, the relative speed of disease arrivals can be accurately predicted using effective distance metrics. In their analysis of the SARS epidemic, their R^2 is 0.973 between the true and predicted arrival rankings based on effective distance, assuming it started in China. Similar accuracy was found for H1N1 assuming it began in Mexico. These findings indicate that the outbreak location of a pandemic has an outsized influence on its eventual spread. When generating a forward-looking analysis or attempting to generalise findings to a future pandemic, an understanding of how results may vary based on the outbreak location of a disease is pertinent. To address the issue, this paper uses compartmental epidemiological models to simulate outbreak scenarios beginning from every available region, reducing the reliance on COVID-19 initial outbreak conditions for the findings on intervention effectiveness.

The outsized influence of the initial outbreak location on the eventual spread

of COVID-19 is pertinent for analyses within Brazil. Coelho et al. (2020) focus on identifying patterns of disease risk from São Paulo and Rio de Janeiro to the rest of Brazil. Using an airline transportation network across 558 micro-regions in the country, they identify those most at risk of early infections through an effective distance analysis. Subsequently, the authors assess the probability of an outbreak in each micro-region across two potential generations of outbreaks. Of the 558 potential micro-regions, the arrival time and disease risk analyses are calculated only from outbreaks beginning in São Paulo or Rio de Janeiro – the two primary transport hubs in the country (Coelho et al., 2020). The arrival time and outbreak probability results vary dramatically based on whether the initial outbreak location is São Paulo or Rio de Janeiro (Coelho et al., 2020). While their analysis provides insightful information on arrival times and disease risks for outbreaks beginning in São Paulo and Rio de Janeiro, the importance of outbreak location constrains its findings to a subset of potential scenarios.

Epidemiological Clustering Analyses

There has been a global effort since the initial outbreak of COVID-19 to reveal underlying patterns of disease risk, whether on COVID-19-specific metrics or socio-demographic vulnerabilities that influence disease risk (Zarikas, Poulopoulos, Gareiou, & Zervas, 2020; Coelho et al., 2020). Unsupervised learning methods, analysing the intrinsic structures within data to uncover hidden patterns or groupings, are of great value to achieve such ends (Guo et al., 2017). Related to COVID-19 metrics, Zarikas et al. (2020) conduct a time-series clustering analysis on countries based on the number of active cases, active cases per population, and active cases per population per area. Their analysis uncovered groups of countries that shared similar patterns, identifying clusters with distinct experiences during the outbreak. Coelho et al. (2020) used k-means clustering to characterise social vulnerability across Brazil, then linked them with a separate outbreak probability analysis. Each of the 558 micro-region was assigned to one of five distinct classes based on its social vulnerability characteristics. The results were merged with an outbreak probability analysis to detect the socially vulnerable regions that were almost certain to have an outbreak. An example of a class deemed most vulnerable to the effects of COVID-19 was in the northern region of Brazil, a predominantly rural region contained within the Amazon, with poor access to education, water, sewage disposal, and electricity (Coelho et al., 2020).

2 Methods

2.1 Data

Cellphone Mobility Data

Through an agreement with In Loco, a Brazilian cellphone analytics company, this paper has access to daily isolation levels for 2737 regions of roughly 740m^2 across the MRSP. Each region is a hexagon of uniform size and shape. In Loco gathers data by partnering with mobile phone applications, and uses software development kits to harvest location data while individuals are using the partnered app (Peixoto et al., 2020). This form of location gathering provides precise geo-coordinates, which are anonymised and aggregated to develop the social-isolation indices. Daily isolation levels across 2737 regions are present from March 1, 2020 to August 19, 2020. For a given region, the proportion of individuals who reside in the region and *stay* in that region on a given day is recorded. This proportional value is used as a proxy for social isolation, recording the extent to which individuals travel outside their residence area. Higher social isolation values indicate that *less* individuals are leaving their residence area, and vice versa for lower social isolation values. The distribution of social isolation regions is presented in Figure 3. The uneven coverage of the MRSP regions is a feature of the data set provided by In Loco, discussed in the limitations section.

COVID-19 Data Set

For this analysis, data on COVID-19 are gathered from the São Paulo State Health SARI hospitalisations database (Secretaria da saúde do estado de São Paulo, 2020). The Brazilian Ministry of Health defines SARIs, which can be caused by COVID-19, as a flu-like syndrome which includes one of: persistent chest pain; laboured breathing; or hypoxia (Li, Sabrina, Pereira et al., 2020). The database records the demographic characteristics, address, whether hospitalisation occured, and health outcome (recovered or death). Hospitalisation is not a pre-requisite for inclusion in the database, and all cases are recorded (Li, Sabrina, Pereira et al., 2020). The same data set is used in Li, Sabrina, Pereira et al.'s (2020) analysis of social and racial inequities of hospitalisation and death across the MRSP. COVID-19-specific data was not available in Brazil at the outset of the pandemic, a challenge faced by researchers assessing its initial spread (Candido et al., 2020). This is largely due to changing notifications and inconsistent access to testing across Brazil (Candido et al., 2020). As done in Candido et al. (2020), the primary metric used to assess transmission

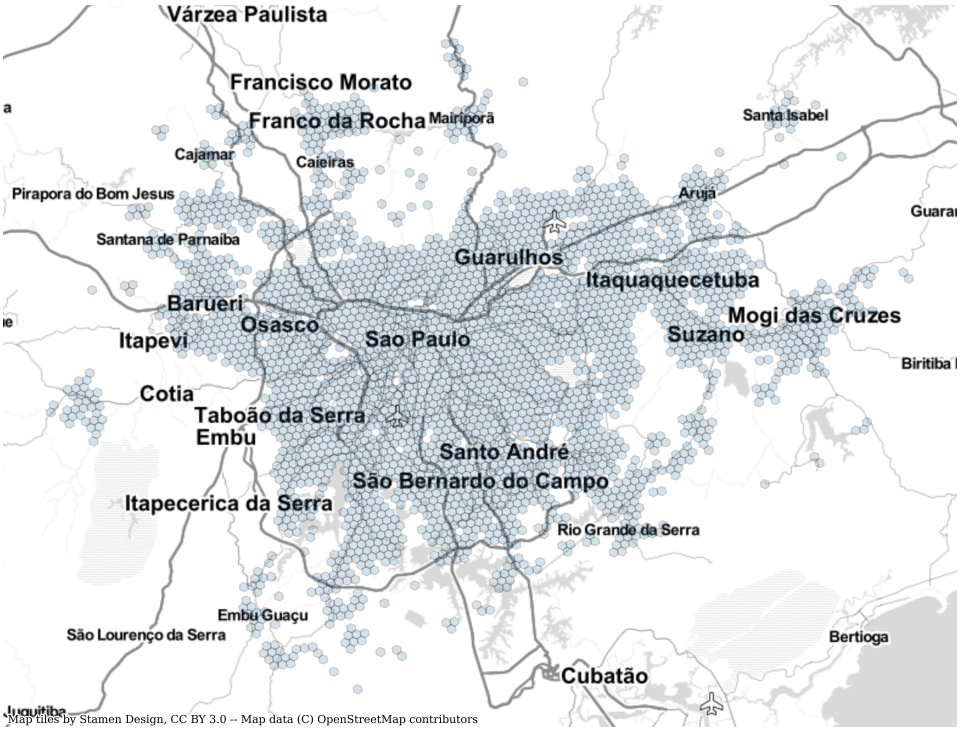


Figure 3: Distribution of all In Loco social isolation hexagons in the MRSP

timelines is the deaths caused by SARIs in the data base, as minor COVID-19 case reporting is more variable as a result of the aforementioned limitations.

Commuting Data Set

The commuting data for the MRSP were gathered from the 2017 MRSP household travel survey, conducted by the São Paulo Metropolitan Transportation Department (METRÔ-SP, 2017). The original data set is a table of survey responses regarding the daily commutes of 86,318 individuals who reside in the MRSP. On average, each individual reports 2.12 daily journeys, leading to a total of 182,994 trip reports. Key information for the reports are the journey origin and destination, along with the travel time. The interviews were conducted across 39 municipalities within the MRSP, divided into 510 research zones for the purposes of the study. Of all the research zones, 66% lie within the main municipality in the MRSP, São Paulo. The survey was designed to be statistically representative across the MRSP, and includes journey and population weights to scale responses by their frequency in the true population. From these weights, the total 2017 commuting flow between

regions and 2017 estimates of populations were calculated.

Census Data Set

This study uses socio-demographic data from the official 2010 Brazilian Census, focused on the census tracts within the MRSP (Instituto Brasileiro de Geografia e Estatística, 2011). Within the state of São Paulo, there are 68,296 tracts, with data included on the total population, racial aggregates, average per capita Brazilian Real per calendar month (R$/pcm), functioning water networks, and other relevant socio-demographic features. The census regions within São Paulo state cover a larger area than both the cellphone mobility and commuting data, which are primarily focused on the MRSP. The population data from the the MRSP commuting survey is more up to date than the 2010 census, therefore it is used in favor of the census data population totals. The census data remains useful for calculating regional income and proportional values for racial variables and access to water networks.

2.2 Data Preprocessing Steps

When assessing intervention effectiveness using cellphone mobility data from 2020, using population and commuting values from 2017 may introduce erroneous estimates that do not represent the MRSP at the outset of the pandemic. This type of processing error, arising from data linkage with inconsistent time periods, is common when merging traditional and Big Data sets (Amaya, Biemer, & Kinyon, 2020) To fix this issue, the first major data preprocessing step in this thesis is to estimate the population and commuting data for 2020 based on MRSP data from prior years.

To integrate the census, COVID-19, and commuting data sets for analysis, a second substantial data preprocessing step is to interpolate them onto the same spatial scale as the cellphone-based social isolation hexagons. This step leverages the mobility data provided by In Loco without further aggregations, maximising its spatial resolution.

Estimating 2020 Populations

To generate 2020 population estimates for each commuting survey region, the geometric growth rate of the population is estimated using records from the 2010 census and an official 2019 statistics report (Instituto Brasileiro de Geografia e Estatística, 2011; Ministério da Economia/Fundação Instituto Brasileiro de Geografia e Estatística, 2019). This method of population extrapolation assumes a stable growth rate year-to-year, where a region's population increases proportionately to

its size (Regmi, 2015). The formula to determine the geometric growth rate (g) is the following:

$$g = \left(\frac{pop_{T_1}}{pop_{T_0}} \right)^{\frac{1}{\triangle T}} \tag{1}$$

where T_1 is the 2019 population, T_0 is the 2010 population, and $\triangle T$ is 9, the number of years between the estimates. Having calculated the growth rate, the population of each commuting region is scaled by this growth rate. As we are estimating the 2020 population values from the 2017 survey data, the following equation is used:

$$pop_{2020} = pop_{2017} * g^{(2020-2017)} \tag{2}$$

The geometric growth rate is to the power of three to represent the three years of growth from 2017 until 2020. The 2020 population estimates for each region are then used to interpolate the hexagon populations through aerial weighting.

Estimating 2020 Commuting Flow

To estimate the 2020 commuting flow for each origin and destination region pair in the 2017 MRSP travel survey, this analysis employs a modified radiation model from Yang, Herrera, Eagle, and González (2014). The radiation model relies on the assumption that the total commuter flux between two regions is based on the number of 'intervening opportunities' between the origin and destination, rather than only their distance (Barbosa et al., 2018; Yang et al., 2014). The original radiation model for estimating and predicting commuting flow – which Yang et al. (2014) expand on – has the following structure:

$$T_{ij} = O_i \frac{n_i n_j}{(n_i + s_{ij})(n_i + n_j + s_{ij})} \tag{3}$$

where O_i is the total volume of trips, and n_i and n_j are the origin and destination population values, respectively. A radius is defined as the distance between i and j, and s_{ij} represents the population within the circle defined by this radius. This s_{ij} value acts as a proxy for the number of opportunities between i and j. The original model is parameter free, not requiring any calibration. This proves problematic when estimating regional commuting flows for an urban agglomeration such as the MRSP, where many regions serve as potential business hubs for commuters (Yang et al., 2014).

To expand on the original radiation model, Yang et al. (2014) derive a radiation model based on a survival analysis framework, and introduce a parameter α which is fit to the observed commuting data:

$$T_{ij} = \lambda \frac{\left[(a_{ij} + n_j)^{\alpha} - a_{ij}^{\alpha}\right] (n_i^{\alpha} + 1)}{\left(a_{ij}^{\alpha} + 1\right) \left[(a_{ij} + n_j)^{\alpha} + 1\right]}. \tag{4}$$

In this model, a_{ij} is equivalent to $n_i + s_{ij}$, which is the sum of the population of region i and *in between* i and j (not including population of j). The variable n_j is the population of region j, and λ is a scaling parameter defined as the ratio of commuters to the population (Yang et al., 2014). Rather than use the geographic distance between i and j to estimate the intervening opportunities, this analysis uses the mean commuting time between the regions, as reported in the commuting survey. Thus, a region is considered in between i and j if its commuting time from region i is less than the commuting from i to j. The population of all regions which match this criteria are used to calculate s_{ij}. The commuting time is better suited to determine intervening opportunities relative to an absolute geographic distance, as it more directly reflects the costs to a commuter deciding to make a journey, and takes into account the topological structure of the commuting network. For example, regions which are closer together physically may require a more costly journey due to traffic or the road network, which commuting time would capture. The 2017 commuting survey is used as input data to estimate the α parameter, optimised by minimising the squared error between the observed and estimated commuting flow. Finally, the 2020 commuting estimates are generated using Yang et al. (2014)'s model with the fitted α parameter and the 2020 population estimates for each zone.

Interpolating Data to Hexagon-Level

While the social isolation regions provide spatially and temporally granular information on the daily proportion of residents leaving a given area, information on which population subgroups are included in each hexagon remain unknown. This problem is shared across the growing body of literature using cellphone mobility data for public health purposes, where anonymity measures by cellphone data providers obscure information on the sample (Grantz et al., 2020). While fundamental selection biases in the mobile phone data are a lingering issue, discussed in the limitations section, traditional data sources can be leveraged to generate population estimates within the hexagons (Aleta, Mart, Bakker, Pastore, & Ajelli, 2020).

Whereas the COVID-19 data is precise to the postal-code level and can be assigned into hexagons, the census and commuting data is constructed of varying

spatial structures which must be mapped to the social isolation hexagons. This process, known as spatial interpolation, is used in geospatial studies to estimate values in unknown area units using values in known geographic units (Comber & Zeng, 2019). Figure 4 depicts the extent to which census regions (blue) overlap with the social isolation hexagon (purple), highlighting the importance of a sound interpolation method to map socio-demographic information.



Figure 4: Illustration of a single hexagon (purple) and its overlap with multiple census regions (blue).

The spatial interpolation method used in this analysis known as aerial weighting, which integrates socio-demographic estimates based on proportional overlap (Comber & Zeng, 2019). This method depends on the assumption of homogeneously distributed characteristics across census and commuting regions, but benefits from transparency and simplicity relative to interpolation methods which rely on auxiliary information (Comber & Zeng, 2019). Each hexagon's overlap with census regions was determined relative to the census region's total area. This proportional overlap area was used to generate a weighted allocation for racial composition, income, functioning water networks, and 2020 estimate populations. For example, if a hexagon covered 50% of a census region with a population of 20, the hexagon would be assigned 10 individuals. Figure 5 displays the distribution of interpolated populations across the hexagons in the MRSP.

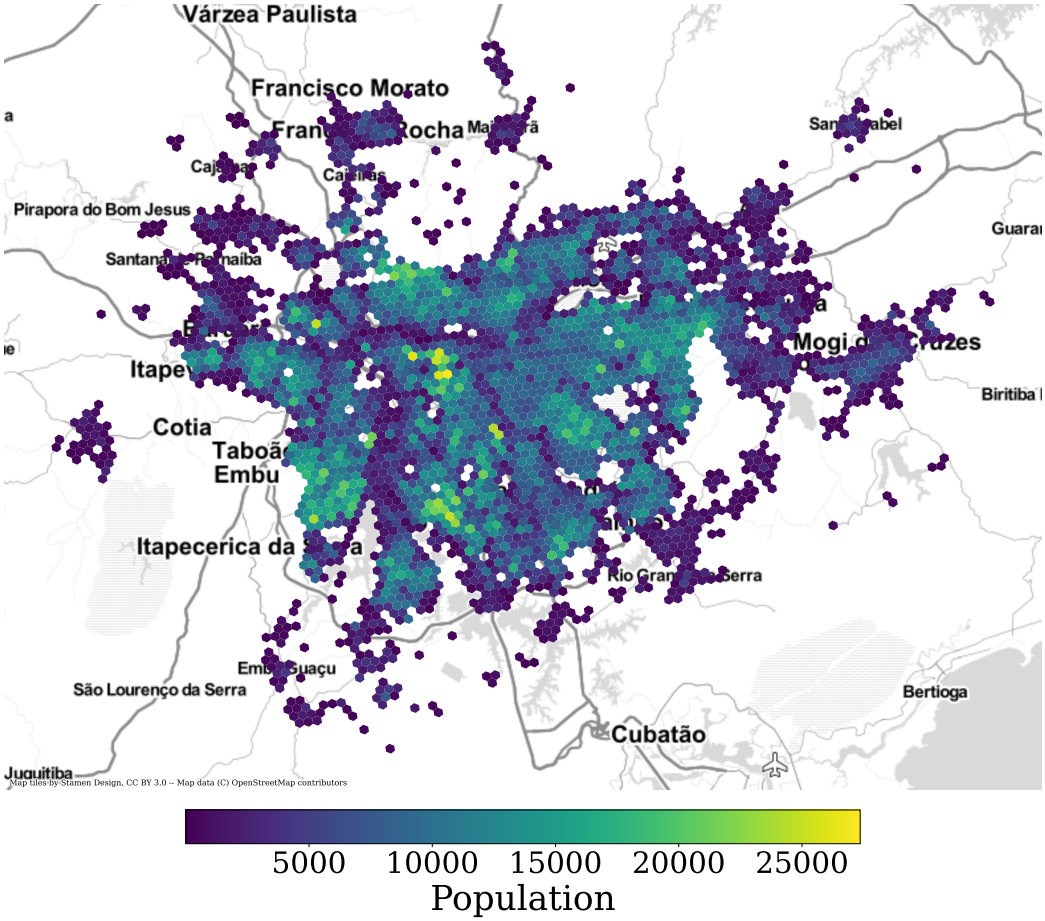


Figure 5: Distribution of interpolated population of each social isolation hexagon, using 2020 estimates. Shows variation in populations across the MRSP.

To interpolate the estimated 2020 commuting network to the isolation hexagons, the homogeneity assumptions of aerial weighting are extended to commuting flow between regions (Jang & Yao, 2011). It is assumed that a hexagon overlapping with a given origin zone has a proportional quantity of outflow to all its targets. Similarly, hexagons overlapping with a given destination zone receive commuter flow from all relevant origin zones proportional to their intersection with that destination zone. An example of the interpolation commuting method is provided in Figure 6.

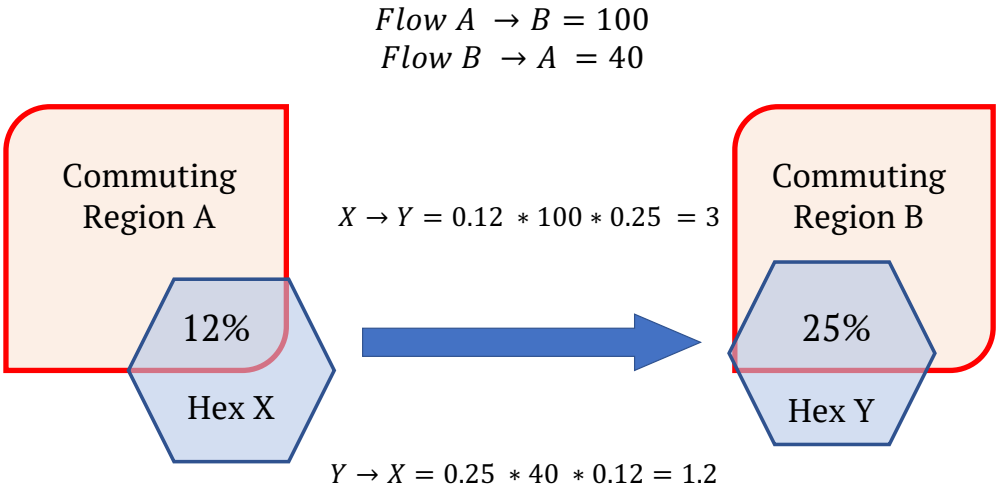


Figure 6: Commuter flow data interpolation strategy from commuting regions to social isolation hexagons. Outflow and inflow are proportional to a hexagon's overlap with a commuting zone. To estimate commuting patterns within a given hexagon, known inflow and outflow between commuting regions A and B are proportionally allocated based on the overlap. For example, Hexagon X overlaps with 12% of Region A, and therefore 12% of Region A's *outflow* is assigned to Hexagon X. Hexagon Y overlaps with 25% of Region B, and therefore 25% of the 12% outflow is assigned to Hexagon Y as *inflow*.

Based on the interpolated commuting network, the median in-degree and out-degree of each region is assessed to measure their centrality. In-degree centrality is the number of edges that directly flow into a region, and out-degree centrality is the number of edges flowing outwards (Hunter, Namee, & Kelleher, 2020). These are highly relevant metrics for an analysis based on disease spread, representing the diversity of regions flowing inwards and outwards, found to be associated with a region's time to infection (Christley et al., 2005). Additionally, the weighted in-degree (total commuters in) and weighted out-degree (total commuters out) is also calculated for each region, shown in Francetic and Munford (2021) to influence the spread of disease.

Across all hexagons, the distributions of commuting, population, race, and income characteristics are summarised in Figure 7. A table of summary statistics is presented in Appendix A.

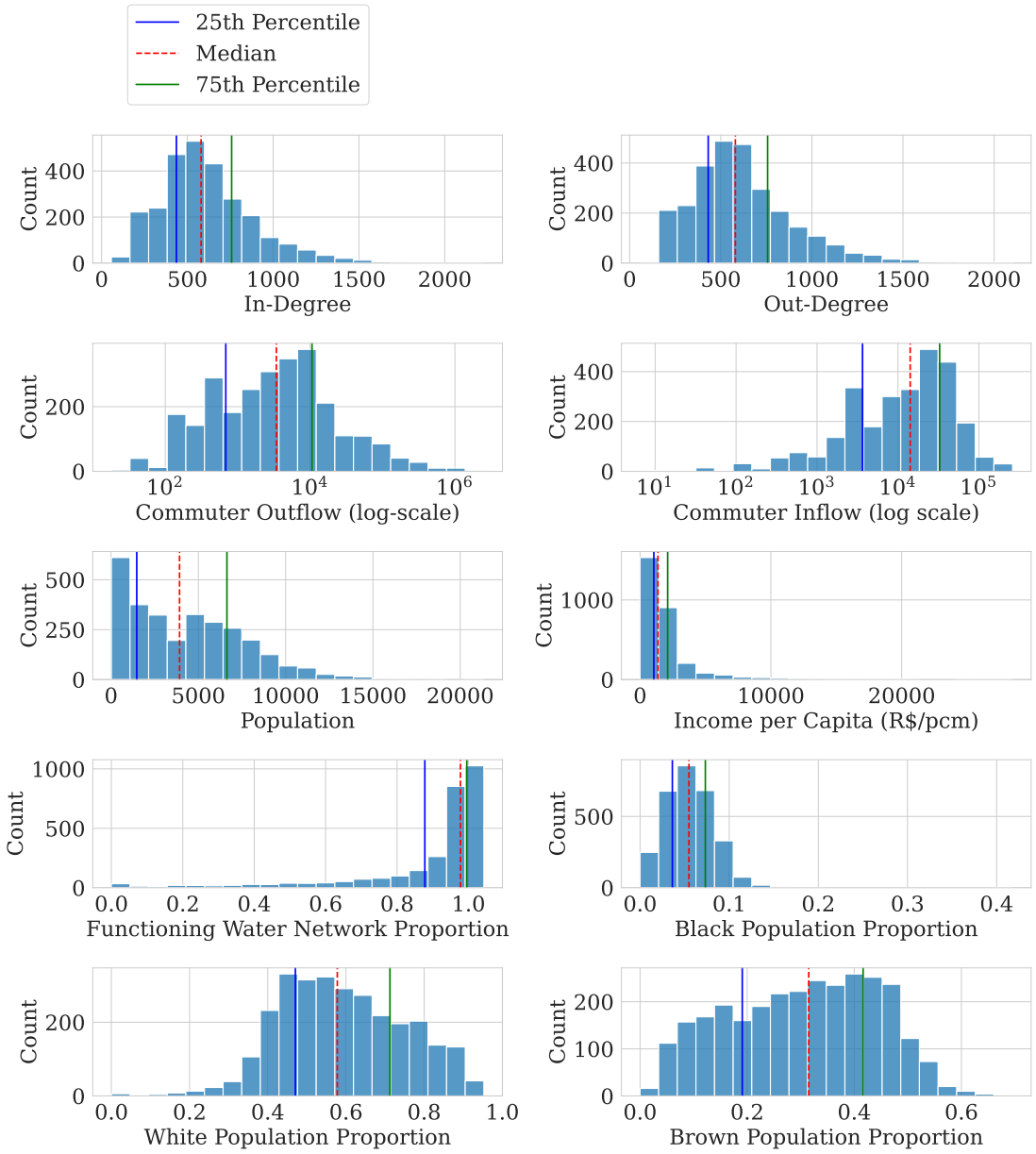


Figure 7: The socio-demographic and centrality distributions of all 2737 social isolation hexagons.

2.3 Infection Delay Model

The inputs to the infection delay model are the effective distances from all pairs of hexagons, calculated under baseline and intervention mobility scenarios. These effective distances are translated into two sets of infection arrival times, whose differences represent the 'infection delay' of an intervention. Subsequently, an algorithm is developed to compute the infection delays of an intervention for a given region over the course of an outbreak, outputting a time series plot defined as the infection delay curve.

Calculating Effective Distance

To calculate the effective distances for the hexagon-scaled commuting network, this analysis uses the 'dominant path' effective distance, used in the disease arrival time analyses of Brockmann and Helbing (2014), Coelho et al. (2020), and Gautreau et al. (2008). Dominant path effective distance measures focus solely on the most probable path of transmission from region i to j. To calculate this value, for every connected origin i and destination j in the network, we calculate the transition rate as the proportion of total commuters beginning in node i who arrive in node j, denoted as $0 \leq P_{ji} \leq 1$ (Brockmann & Helbing, 2014). The effective distance between nodes i and j is calculated as:

$$d_{ji} = 1 - \log P_{ji}, \tag{5}$$

which is used as an edge-weight for every pair of i, j nodes, as shown in Brockmann and Helbing (2014). These edge weights, greater than or equal to one, are used in a weighted shortest path analysis to determine the dominant path effective distance between every pair of regions. With d_{ji} calculated for all edges in the network, the dominant path between i and j is chosen as the path which minimises the sum of effective distance edge weights between them. Finally, the dominant path effective distance between two nodes (D_{ji}) is calculated as the sum of the effective distances along the determined shortest path. This basic dominant path effective distance can be used to detect rankings of arrival times for a given outbreak location (Brockmann & Helbing, 2014).

The traditional dominant path effective distance model is solely based on the mobility network, and cannot be influenced by changing epidemiological parameters or mobility rates in a commuting network. To add epidemiological and mobility-based parameters, useful for a comparative analysis, the effective distance formula is altered to

$$d_{ji} = \log\left(\frac{\beta - \mu}{1 - \kappa} - \lambda\right) - \log(P_{ji}), \tag{6}$$

shown in Iannelli, Koher, Brockmann, Hövel, and Sokolov (2017), where β and μ are the infection and recovery rate, κ is the mobility reduction, and λ is the Euler-Mascheroni constant . The reproductive number R_0 is chosen to be 2.9, based on W. M. de Souza et al. (2020)'s epidemiological characterisation of the MRSP early in the pandemic. The infectious period is chosen to be 9.2 from Pinto Neto et al. (2021)'s mathematical modelling of COVID-19 in Brazil. The infection rate is thus $R_0/\textit{infectious period} = 2.9/9.2$, and the recovery rate is given by $1/\textit{infectious period}$=1/9.2. It is important to note that the transition rate P_{ji} calculation is unaltered from the traditional model. As the mobility reduction κ rises, d_{ji} decreases, indicating the i and j are effectively farther. This modified equation does not impact the rankings of arrival times from a given outbreak location, but allows for varying effective distance calculations under different mobility scenarios. Similarly to the traditional model, for every potential outbreak and target region in the network, the dominant path effective distance is generated from the weighted shortest path analysis, generating an 2737×2737 matrix of effective distances. This method is able to calculate effective distances between nodes irrespective of whether they are directly or indirectly connected.

Two 2737×2737 matrices of effective distances are calculated for every potential origin i and destination j, under the following commuter flow scenarios:

1. No traffic reduction

2. Reduction in traffic based on hexagonal isolation changes

The first commuter flow scenario assumes no interventions ($k = 0$), where arrival times are calculated using the baseline commuting information. The second scenario assumes that regions reduce their mobility by the same amount as observed in the pandemic, through leveraging the cellphone social isolation information. To determine the extent of the mobility reduction for each region, the marginal change in social isolation from pre-pandemic to mid-pandemic is calculated. The initial isolation value for each hexagon is calculated as the mean across March 1 to March 15, the two weeks leading up to the MRSP's lockdown. The pandemic isolation value for each hexagon is calculated as mean from March 16 to March 30 2020, capturing the initial regional responses to lockdown measures. After determining the marginal change in real isolation for each origin hexagon, the effective distance calculation becomes:

$$d_{ji}^{intervention} = \log\left(\frac{\beta - \mu}{1 - \kappa_i^{marginal\ change}} - \lambda\right) - \log(P_{ji}) \tag{7}$$

This representation of effective distance is used to approximate how rapidly a disease would spread from hexagon i to j given the observed change in pandemic isolation for region i. The adjustment of the κ term to $\kappa_i^{marginal\ change}$ is a novel contribution of the study, allowing the analysis to capture heterogeneous changes in mobility based on cellphone mobility data. Incorporating regional differences in mobility is especially important given racial and income-related inequalities in social isolation abilities (Li, Sabrina, Pereira et al., 2020).

Infection Delay of Intervention

To generate an estimation of arrival times based on the effective distances, this paper employs the methods used in Iannelli et al. (2017), dividing the effective distance by the effective velocity, defined as $V^{EF} \approx \beta - \mu$, where β is the infection rate and μ is the recovery rate. The arrival time for a disease to arrive in location j from location i, including both the sum of shortest path effective distances D_{ji} and velocity is thus:

$$T_{ji} = \frac{D_{ji}(P)}{V^{EF}} \tag{8}$$

Having generated the arrival times under both scenarios for every i, j combination, the infection delay by an intervention for an introductory case arriving from origin i to destination j is calculated as:

$$ID_{ji} = T_{ji}^{intervention} - T_{ji}^{baseline} \tag{9}$$

The infection delay (ID_{ji}) values are calculated for every pair of nodes, generating a 2737×2737 matrix where each i, j value represents the additional time to a case arriving to node j from node i given a reduction a traffic proportional to node i's real mobility change.

Using known changes in mobility to understand intervention effectiveness takes into account the inequality in regional responses, and allows intervention scenarios to mimic the real capacities of regions to isolate and adhere to policy guidelines. Having the arrival times in $T_{ji}^{intervention}$ reflecting the real mobility changes allows for an infection delay analysis which better captures the lived experience of each of the 2737 regions in determining the relative benefits from early interventions.

From the MRSP's first case of COVID-19 to its widespread presence, this anal-

ysis determines the time 'added' until a region's first case (infection delay) by an intervention at every hypothetical time t, assuming no intervention before t. At time $t = 0$, only the initial outbreak location i_0 has the disease, and each hexagon's infection delay by an intervention is $ID^0_{ji_0}$, representing the change in intervention arrival time relative to the baseline arrival time from i_0 to j. For every $t \geq 1$, each hexagon's infection delay value is determined based on the currently infected regions. To calculate this value, for every region j and discrete time step t, the following algorithm is developed:

1. Determine all infected regions at time t
2. Calculate the infection delay of an intervention across all currently infected regions relative to destination j
3. Select the minimum infection delay value

The IDM generates a time-series plot for each region, known as its infection delay curve. An example plot is presented in Figure 8, for a given hexagon A and outbreak location B. There are two factors that interact to create the slope of the infection delay curve: (1) the effective distance of infected regions to the region at risk; (2) the degree of mobility reduction of infected regions. The outbreaks used in this analysis will be simulations calculated from a compartmental epidemiological model, detailed in Section 2.4.

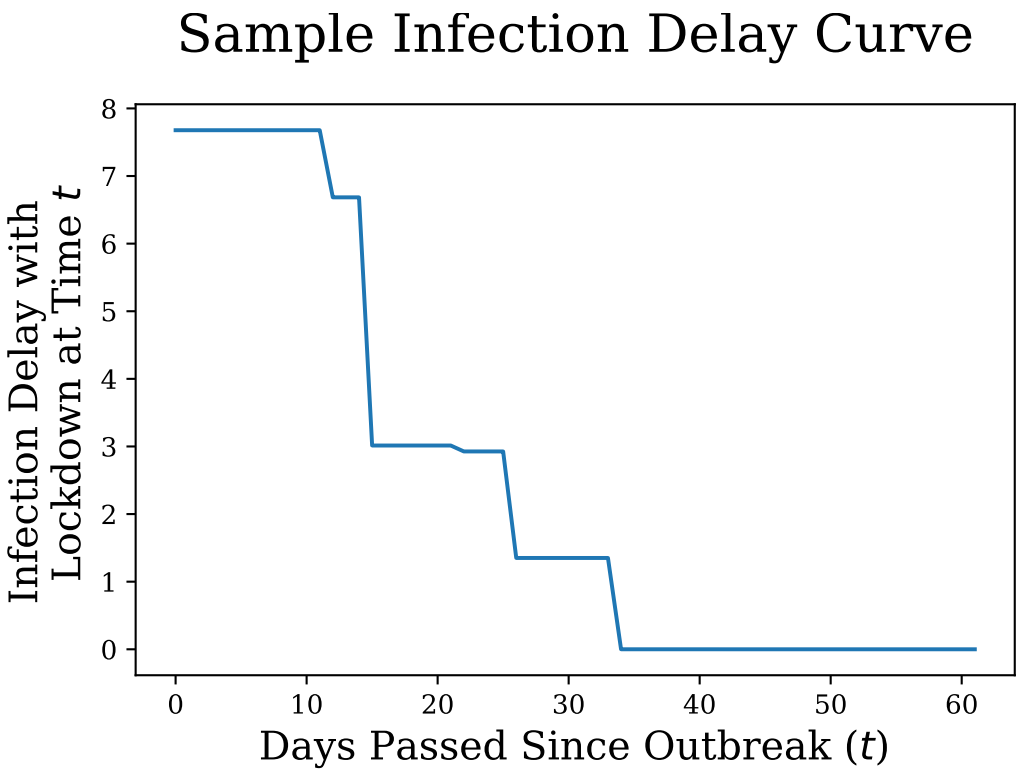


Figure 8: Hypothetical infection delay curve for region-at-risk A caused by a lockdown, following an outbreak beginning in region B. At time $t = 0$, location B would be the only infected region – as the outbreak location. At this time, a lockdown would allow region A to gain approximately 7.5 days (y-axis) until its first case of COVID-19. If the disease was to spread unmitigated until time $t = 20$ days, a lockdown would provide a gain of only 3 days before region A's first case. At the 34-day mark following an outbreak in region B, without any intervention, region A would already be infected. Thus, a lockdown intervention at this point would have no ability to delay the onset of infection, with a y-axis value of 0.

2.4 Commuter-Oriented Epidemic Simulations

To generalise the findings of the infection delay analysis to outbreak scenarios other than those observed during COVID-19, epidemic outbreaks are simulated beginning in each of the 2737 hexagon regions in the MRSP. This paper employs the commuter-oriented susceptible-infected-removed (SIR) model, shown in Schlosser et al. (2021). While the original model is used to simulate the spread of COVID-19 in 401 German counties, this analysis uses the 2737 social isolation hexagons, providing their interpolated 2020 estimates of populations and commuting flows. Every member of each population is divided into three possible states: susceptible (S); infected (I); and recovered (R), where $N_i^{pop} = S_i + I_i + R_i$.

Being a metapopulation model, every individual is assigned a 'home' region i and and a work region j. To assign individuals with home location i to a work

region, a commuter probability matrix is constructed by finding the proportion of individuals who commute to j from i (Schlosser et al., 2021). These probabilities are then multiplied by population values so that each individual commutes to a work location proportionally to the commuter flow defined by the mobility matrix. Every home population is divided into subgroups with the same j work destinations, defined as N_{ji}^{pop}, where $N_i^{pop} = \sum_j N_{ji}^{pop}$.

When initiating an outbreak, every individual in each N_{ji} subgroup is assigned as susceptible, such that $S_{ji} = N_{ji}^{pop}$, excluding the outbreak location i_0 (Schlosser et al., 2021). At time $t = 0$, X infected individuals are distributed across the N_{ji_0} subpopulations, proportionally to their size relative to the total outbreak location, $N_{i_0}^{pop}$. It is assumed that in each time period t, individuals spend 50% their time in the home location and 50% at the work location. Susceptible individuals S_{ji} can be infected during the home phase by infected residents of location i, or in the work phase by individuals who commute to region j.

The rate of change for susceptible individuals S_{ji} (commuting to location j from location i) to the infected compartment is given by:

$$\frac{\mathrm{d}S_{ji}}{\mathrm{d}t} = -S_{ji}\left(\lambda_i^{\text{home}} + \lambda_j^{\text{work}}\right) = -S_{ji}\left(\frac{\beta}{2}\frac{\sum_k^m I_{ki}}{\sum_k^m N_{ki}^{(\text{pop})}} + \frac{\beta}{2}\frac{\sum_k^m I_{jk}}{\sum_k^m N_{jk}^{(\text{pop})}}\right) \tag{10}$$

where λ_i^{home} is the infection force at home defined by the proportion of all infected individuals at home in i relative to the total number of individuals whose home location is i (Schlosser et al., 2021). The infection force from commuting to work is λ_i^{work}, defined by all the infected individuals commuting to region j divided by the population of commuters to region j from all regions. The β parameter is the infection rate of the disease.

Whereas the rate of change for susceptible individuals is negative as individuals become infected, this same mechanism leads to a positive rate of change in infected individuals, given by:

$$\frac{\mathrm{d}I_{ji}}{\mathrm{d}t} = -\mu I_{ji} + S_{ji}\left(\lambda_i^{\text{home}} + \lambda_j^{\text{work}}\right) \tag{11}$$

Where $-\mu$ is the recovery rate. The change from infected to recovered is simply given as:

$$\frac{\mathrm{d}R_{ji}}{\mathrm{d}t} = -\mu I_{ji} \tag{12}$$

A binomial sampling algorithm is used to used to numerically simulate the epidemic

(Schlosser et al., 2021). From a time interval $[t,\ t + \triangle t]$, the probability that an individual in the S_{ji} compartment becomes infected by the force of transmission $\lambda_{ji} = \lambda_j^{\text{home}} + \lambda_j^{\text{work}}$ is given by:

$$P\left(\Delta t; \lambda_{ji}\right) = 1 - \mathrm{e}^{-\lambda_{ji}\Delta t} \tag{13}$$

The individuals in compartment S_{ji} that become infected during this interval are drawn from the binomial distribution with probability $P\left(\Delta t; \lambda_{ji}\right)$ as

$$\Delta\left(S_{ji} \to I_{ji}\right) \sim \text{Binom}\left(S_{ji}(t), P\left(\Delta t; \lambda_{ji}\right)\right) \tag{14}$$

and the individuals that recover are given by:

$$\Delta\left(I_{ji} \to R_{ji}\right) \sim \text{Binom}\left(I_{ji}(t), P(\Delta t; \mu)\right) \tag{15}$$

Key epidemiological parameters for the model are the reproductive number (R_0) and the recovery rate (μ). The R_0 value and μ are the same values used in Section 2.3, from W. M. de Souza et al. (2020) and Pinto Neto et al. (2021), respectively. The simulation runs beginning from $t = 0$, with time-steps $\triangle t = 1$, up to a maximum of $t = 100$. As each of 2737 simulations progress, the arrival times of the disease to each hexagon is recorded, which is the necessary component for the infection delay analysis.

To create the generalised model of intervention effectiveness, 2737 epidemics are simulated using the SIR model, beginning in each hexagon. For each outbreak scenario, the calculated arrival times are used in conjunction with the IDM to generate infection delay curves for every region, as described in Section 2.3. The end result is 2736 infection delay curves for every hexagon (excluding its own outbreak), each one encapsulating the infection delay to the first case by an intervention at every time t. Having reached this stage, we can address Research Question 1 and Research Question 2, reiterated below:

- Research Question 1 : When *at risk of infection*, how do regions whose first case is delayed the most and least by a lockdown differ on socio-demographic characteristics?

- Research Question 2 : What are the socio-demographic characteristics of the *initial* outbreak locations where lockdowns lead to disproportionately high and low infection delays for the rest of the MRSP regions?

2.5 Effective Distance and Arrival Time Validation

The first result of the paper is a validation of the relationship between effective distance metrics and disease arrival times in the hexagon-interpolated commuting network, to confirm its suitability for the infection delay analysis. Effective distance is indicative of arrival times on networks of varying scales, from regional food transport networks, shown in Brockmann and Helbing (2014), to global airport networks, assessed in Gautreau et al. (2008). The methods in this paper require regions with a larger effective distance from a given outbreak location to have later arrival times, and vice versa for regions with smaller effective distances. Effective distance measures capture the importance of disproportionately dominant transport routes, which are critical paths of disease spread. This allows them to efficiently detect most probable introductory paths despite underlying network complexity and numerous possible paths between two nodes (Brockmann & Helbing, 2014). Brockmann and Helbing (2014) describe how geographically complex transport routes evolve into simple infection waves, allowing the detection of propagation speeds and arrival times based solely on mobility networks.

To conduct this validation, the SARI database from the MRSP is required, with postal codes mapped into the hexagons. The analysis first establishes an estimated outbreak location in the MRSP using the place of the first SARI death at least a week after the first case on February 25, 2020. COVID-19 deaths are used due to more consistent reporting and greater robustness, as is done in the methods of Candido et al. (2020). Each hexagon's COVID-19 introduction date is the day of the first detected SARI death in each hexagon after the initial outbreak. After detecting the first COVID-19 case in each hexagon, the time to a first infection is calculated as the difference between the initial outbreak date and each hexagon's first infection date. The dominant path effective distances for each j hexagon from the known outbreak location i are calculated using the process described in Section 2.3. Given that the regions which received COVID-19 earlier in the MRSP were in the more populated municipalities, this variable is controlled for in the regression analysis. Absolute commuting flows are not included as a control variable due to a lack of independence with the effective distance metrics. Two models are created, with and without a control for population. In Model 1, the dependent variable is the time needed to each region's first infection, and the independent variable is each hexagon's effective distance from the outbreak location, shown in the following equation:

$$\text{ArrivalTime}_j = \beta_0 + \beta_1 * \text{EffectiveDistance}_{ji} \tag{16}$$

In Model 2, population is added as an additional control, generating an equation of the following form:

$$\begin{aligned} \text{ArrivalTime}_j = \beta_0 + \beta_1 * \text{EffectiveDistance}_{ji} \\ + \beta_2 * \text{Population}_j \end{aligned} \tag{17}$$

A confirmed association between a hexagon's COVID-19 arrival time and effective distance from the outbreak hexagon serves as a justification for the use of effective distance for the subsequent analyses, and as a validation of the commuting data interpolation.

2.6 Clustering Early Lockdown Effectiveness using K-Means

In this analysis, each of the 2737 regions-at-risk has 2736 infection delay curves, one for every outbreak scenario, not including its own. The median infection delay curve is determined for every region, representing the average delay to infection following lockdown at every time t. To discern patterns of intervention effectiveness and answer Question 1, this analysis employs k-means clustering across the median infection delay curves of all hexagons. This process is presented visually in Figure 9, using only 100 outbreak locations for illustrative purposes. K-means clustering is an unsupervised learning method used to divide data into distinct groups based on their similarity (Singh, Yadav, & Rana, 2013). The algorithm (1) randomly assigns a set of k initialisation cluster centers, (2) calculates the Euclidean distance from all data points to all cluster centers, using the following equation:

$$\text{Dist}_{XY} = \sqrt{\sum_{k=1}^{m} (X_{ik} - X_{jk})^2}, \tag{18}$$

(3) assigns data points to their nearest cluster centers, and (4) generates new cluster centers based on each cluster's centroid (Singh et al., 2013). Steps 2 through 4 are iteratively repeated until no data points are reassigned, and all groups are maximally similar based on Euclidean distance. This method is chosen for both its computational efficiency and interpretability (Singh et al., 2013).

Each cluster and their average infection delay curves will be characterised by their shape and magnitudes, and assessed based on their socio-demographic characteristics. Socio-demographic information included for each cluster is median population size, income, racial composition, and functioning water network access. The centrality metrics included will be in-degree, out-degree, weighted in-degree (total commuters in) and weighted out-degree (total commuters out).

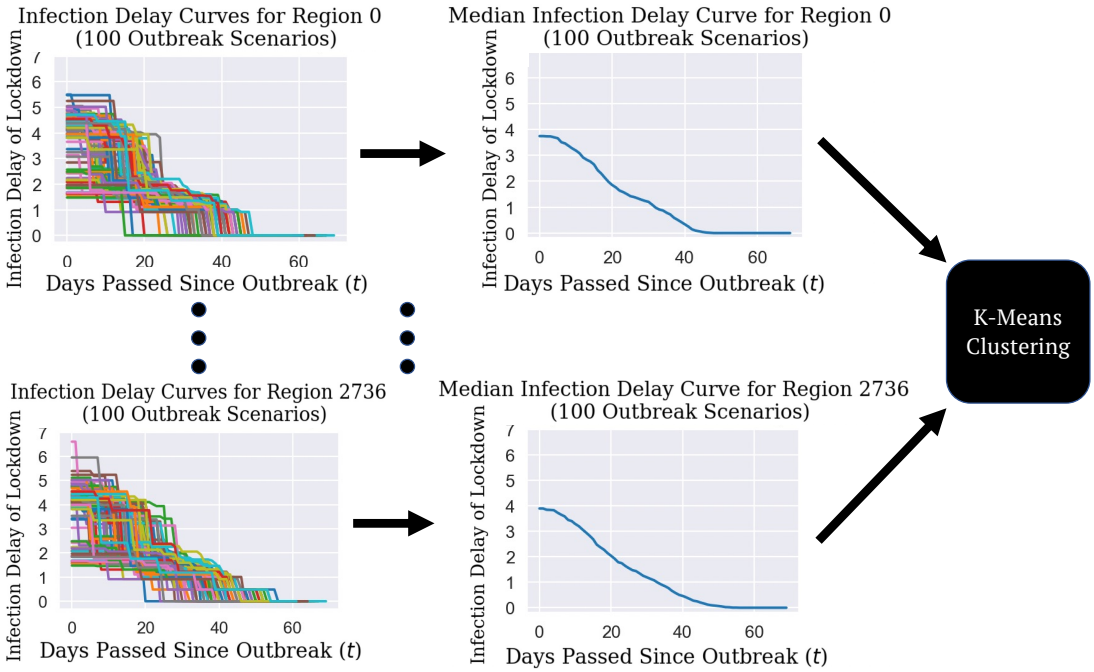


Figure 9: Illustrative example of infection delay clustering pipeline, using only 100 outbreaks for visualisation (real analysis uses 2736 regions of outbreak for each of the 2737 regions-at-risk). From left to right: (1) the infection delay curves of all regions-at-risk are calculated for each outbreak location (shown for Region #0 and #2736); (2) the median of those curves are taken at every time t to create a general characterisation of lockdown effectiveness in the region-at-risk; (3) all median infection delay curves are clustered using k-means, and analysed based on socio-demographic and centrality characteristics.

For each cluster the median infection delay curve will be generated to generally characterise the cluster. Medians are chosen for their greater robustness to outliers, as a mean infection delay is more sensitive to outliers in the cluster (Leys, Ley, Klein, Bernard, & Licata, 2013). Key metrics will be calculated for each cluster, such as the number of days that can be saved from an immediate intervention. Additional assessments include the number of days that can pass since an outbreak before a lockdown can save $[8, 7, ..1]$ days, along with the slope of the curve, broadly indicating the speed of the diminishing returns to an intervention.

2.7 Outlying Outbreak Region Analysis

Research Question 2 is addressed by determining the *initial* outbreak locations where lockdowns lead to disproportionately high and low infection delay values for the rest of the MRSP regions. This analysis is focused on early intervention effectiveness for the 2737 simulated epidemics, and is conducted by looking at the socio-demographic and centrality characteristics of outbreak locations that lead to disproportionately high and low infection delay values (at $t = 0$) for non-infected regions. These insights can be valuable for policy-makers to understand how characteristics of the outbreak location influence the benefits of early lockdowns across the MRSP.

Given a region at risk with 2736 infection delay curves, a single curve is deemed an outlying outbreak region if, at time $t = 0$, it differs from the median of the region's infection delay curves by more than one standard deviation. These outlying outbreak regions are calculated for every region's 2736 infection delay curves, not including its own outbreak curve. Their median demographic variables are incorporated to provide insights on the traits of these outbreak locations. This process is illustrated in Figure 10, using 100 outbreaks for each region-at-risk rather than 2736. For every region, the median characteristics of the positive and negative outliers are calculated, generating two distributions. A two-sample t-test is performed on each demographic and centrality metric to assess the probability that the positive and negative outliers have the same expected values.

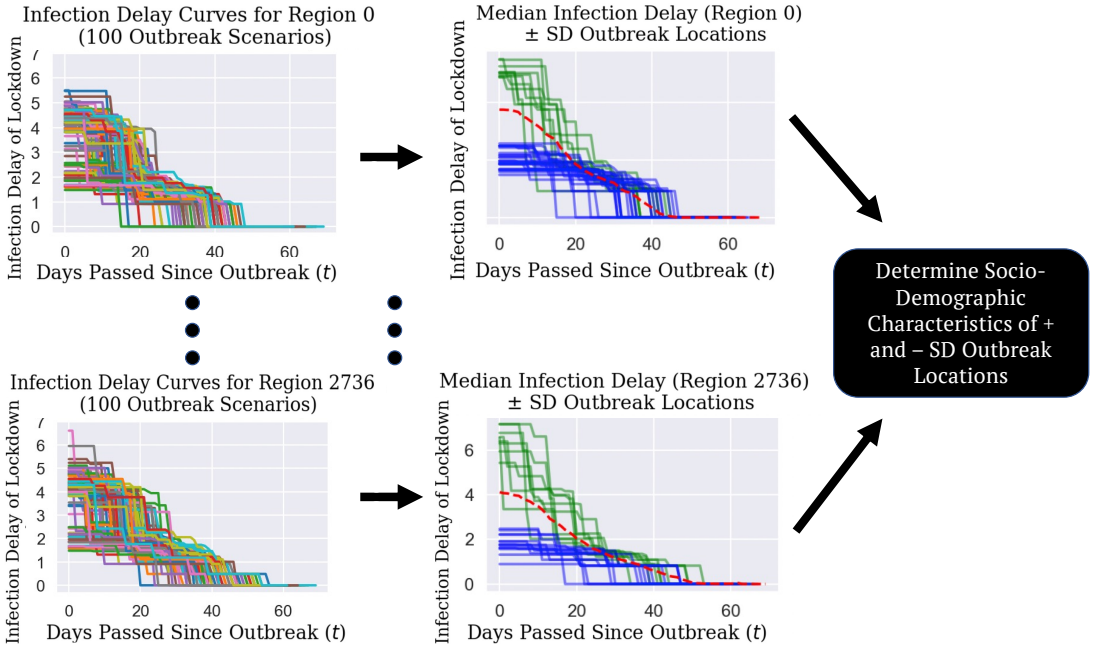


Figure 10: Illustrative example of outlying outbreak region analysis (sample of 100 outbreaks). From left to right: (1) the infection delay curves of all regions-at-risk are calculated for each outbreak location (shown for Region #0 and #2736); (2) The median infection delay curve is determined for each region-at-risk, and used as a cutoff. The curves that are above (+ 1 SD, green) and below (- 1 SD, blue) the median at time $t = 0$ are identified for each region-at-risk and are classified as positive or negative outliers, respectively. For a given region-at-risk, its positive outliers lead to disproportionately large slowdowns, and its negative outliers lead to disproportionately small slowdowns; (3) The socio-demographic and centrality characteristics of both positive and negative outliers are aggregated and presented in the Results section.

3 Results

This section provides the results of the effective distance validation on disease arrival times, using the hexagon interpolated commuting network and the SARI datebase from the MRSP in 2020. Subsequently, the k-means clusters of infection delay curves will be presented and analysed based on centrality and socio-demographic characteristics. Finally, the analysis will assess regions of outbreak where infections can be most and least delayed across the MRSP by early lockdown measures.

3.1 Effective Distance and Arrival Time Validation

The results of the multiple linear regression to determine the relationship between effective distance and arrival times are presented in Table 1. In Model 1 (Equation 16), only regressing on effective distance, there exists a positive relationship between a region's effective distance from the outbreak location and the number of days until its first case, significant at the $p < 0.01$ level. The adjusted R^2 value of 0.38 conveys that 38% of the variation in arrival times are explained by the linear model. The standard error of the effective distance coefficient, representing the estimated standard deviation of the coefficient, is 0.075. Using the standard error to calculate the coefficient's 95% confidence interval, both sides are above 0.0. In Model 2 (Equation 17), where population is controlled for, the effective distance's relationship with arrival time remains positive at the $p < 0.01$ level, and the population has a statistically significant ($p < 0.01$) negative relationship with arrival times. Adding regional populations as a control variable, the adjusted R^2 is 0.457. In both models, there is a positive and statistically significant relationship between between the effective distance from the estimated MRSP outbreak hexagon and a region's estimated COVID-19 arrival time.

	Dependent variable: Disease Arrival Time	
Model	(1)	(2)
Effective Distance	2.083***	3.100***
	(0.053)	(0.073)
Population		-0.003***
		(0.000)
R^2	0.381	0.457
Adjusted R^2	0.380	0.457
Residual Std. Error	28.189 (df = 2518)	26.389(df = 2517)
F Statistic	1547.322*** (df = 1.0; 2518.0)	1061.044*** (df = 2.0; 2517.0)

*p<0.1; **p<0.05; ***p<0.01

Table 1: Multiple linear regression presenting relationship between effective distance and COVID-19 arrival time in MRSP – additional control for population.

3.2 Clustering Infection Delays in Regions-at-Risk

To answer Research Question 1 regarding the socio-demographic characteristics of regions-at-risk that respond similarly to lockdowns, each k-means cluster will be

systematically analysed based on median infection delay curves and their median demographic characteristics. The outbreaks used to calculate each curve are generated from the commuter-oriented SIR model described in Section 2.4. An important feature to explore when analysing each cluster's infection delay curve are the delays caused by an intervention at the initial onset of a disease ($t = 0$). The latest intervention day for clusters to gain $[8, .., 1]$ days is also provided. Additionally, the slope of a line of best fit is presented to depict the speed that interventions lose effectiveness in delaying infection arrivals as an outbreak develops. The median curve characteristics are presented in Table 2, and the median demographic characteristics of each cluster are presented in Table 3, along with their percentile ranks compared to all the MRSP regions. Figure 11 depicts the geographic distribution of the three clusters across the MRSP. In each cluster's subsection visualisation, blue lines represent the median curves of each region in the cluster, and the dotted red line represents the median infection delay curve across the entire cluster (Figures 12, 14, 16).

Cluster	A	B	C
N	550	1179	1008
Infection Delay at $t = 0$	8.82	8.17	7.08
Latest Lockdown to Delay Infection by 8 Days	7	3	N/A
Latest Lockdown to Delay Infection by 7 Days	10	7	2
Latest Lockdown to Delay Infection by 6 Days	13	10	7
Latest Lockdown to Delay Infection by 5 Days	16	13	10
Latest Lockdown to Delay Infection by 4 Days	18	15	13
Latest Lockdown to Delay Infection by 3 Days	22	19	17
Latest Lockdown to Delay Infection by 2 Days	28	25	23
Latest Lockdown to Delay Infection by 1 Days	34	31	29
Slope of Line of Best Fit	-0.22	-0.21	-0.18

Table 2: Key characteristics of each cluster's median infection delay curve. Latest lockdown represented in days since outbreak.

Cluster	A	B	C
In-Degree Centrality	366 (16.3)	567.00 (47.5)	788.00 (77.6)
In-Degree Centrality	362 (16.1)	566.00 (47.6)	796.00 (78.5)
Commuter Out	1095.86 (30.3)	3138.16 (48.1)	7360.12 (66.3)
Commuter In	7892.3 (36.2)	13807.57 (49.3)	21343.39 (60.4)
Population	2890.27 (41.0)	4468.21 (53.1)	4565.34 (54.4)
Income (R$/pcm)	1172.75 (36.5)	1331.55 (48.3)	1729.02 (65.5)
Water Propor.	0.97 (43.5)	0.98 (52.0)	0.98 (55.9)
Black Propor.	0.06 (58.8)	0.06 (54.1)	0.05 (42.2)
White Propor.	0.53 (39.6)	0.56 (46.5)	0.64 (63.2)
Brown Propor.	0.36 (61.0)	0.33 (52.9)	0.25 (35.5)

Table 3: Median and MRSP-wide percentile ranks of centrality and demographic characteristics for each cluster

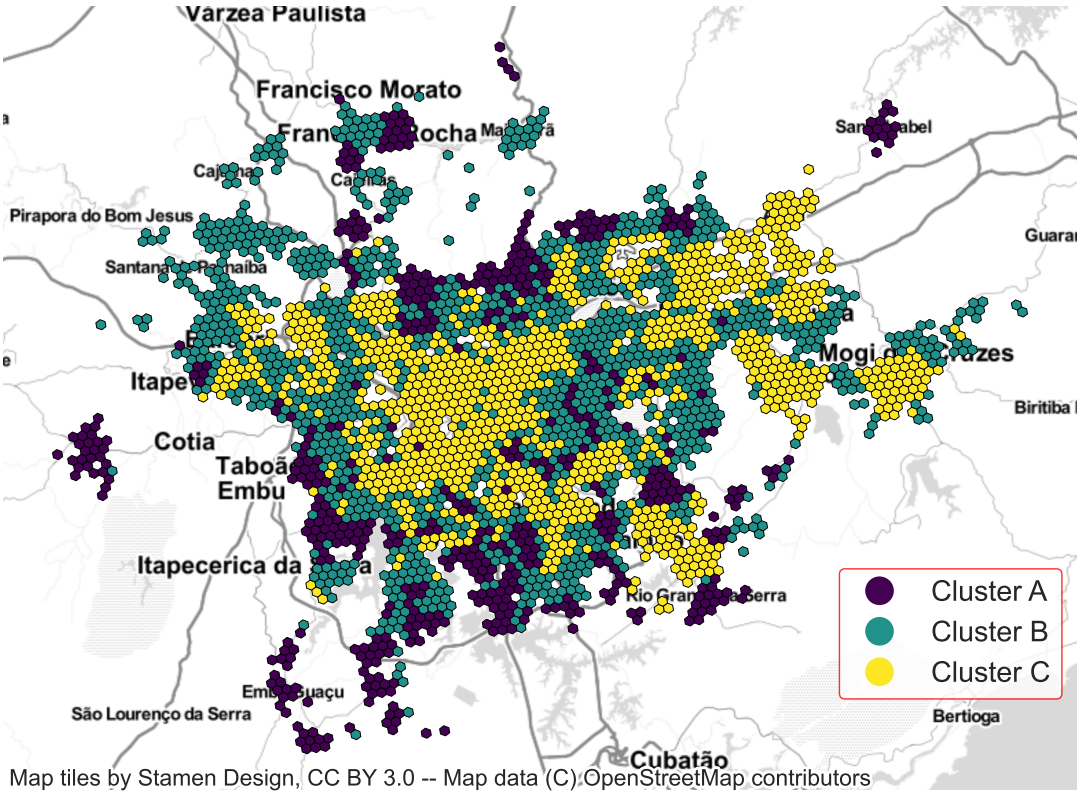


Figure 11: Spatial distribution of all 3 clusters. K-means clustering reveals Cluster A (purple) as the most peripheral, Cluster B (turquoise) in the outer or mid-central regions, and Cluster C (yellow) with the highest central concentration.

Cluster A

For the 550 regions in Cluster A, an intervention immediately after the outbreak of a disease ($t = 0$) would lead to the greatest infection delay, on average, relative to all clusters (Table 2). At this point in time, a mobility reduction as seen during the COVID-19 pandemic would save on average 8.82 days until their first case, taking into account all possible outbreak locations. On average, to delay the onset of disease by at least 8 days, an intervention would be required within the first 7 days after an outbreak. To delay the first infection of a disease by 5 days, an intervention would be required on or before the 16th day after an outbreak. The slope of a linear regression between the days gained of an intervention and the time after the initial outbreak is -0.22. This statistic broadly represents the speed at which a day passing since the outbreak reduces the infection delay of an intervention. The median infection delay curve of Cluster A is presented in Figure 12.

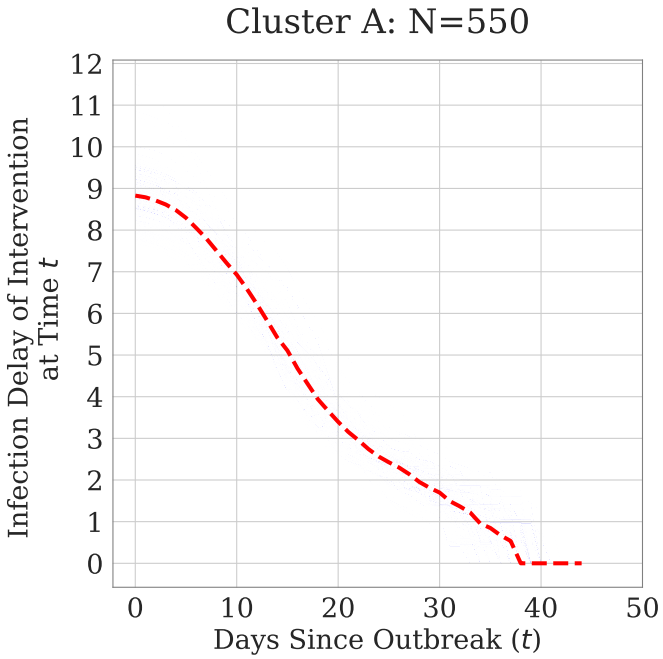


Figure 12: Cluster A infection delay curves (median in red). At time $t = 0$, Cluster A has the highest median infection delay curve of all clusters, on average gaining 8.82 days until its first case following a lockdown.

Cluster A has a median in-degree centrality of 366, and out-degree centrality of 362, both at the 16th percentile of all regions across the MRSP (Table 3). Its median commuting outflow is 1095.86, at the 30th percentile, and its median commuter inflow is 7892.36, at the 36th percentile. The median population of the cluster is 2890.27, at the 41st percentile of all interpolated regions in the MRSP. The median income per capita of Cluster A is 1172.75, at the 36th percentile rank. The black

population of 6% is higher than the median, ranked at the 58th percentile. The white proportion of the population is 53%, at the 39th percentile across the MRSP. The brown percentage is 36%, at the 61st percentile. The proportion of individuals with an active water network is 95%, ranked at the 38th percentile. The socio-demographic and centrality characteristics of Cluster A are plotted in Figure 13. The geographic distribution of regions in Cluster A are presented in Figure 11. They are predominantly on the outer edges of the In Loco coverage, with very little penetration to the outer-central and central MRSP area.

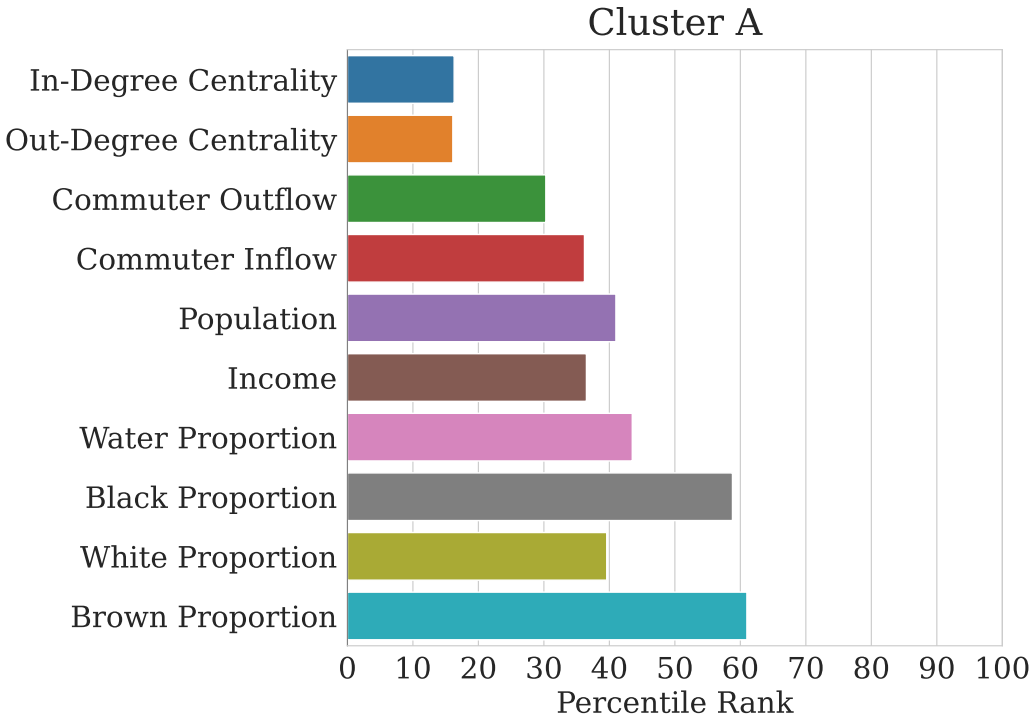


Figure 13: Cluster A: Percentile rank of median centrality and socio-demographic characteristics, relative to all interpolated MRSP regions. In-and-out-degree centrality has the lowest relative rankings, and black and brown population proportions have the highest relative rankings.

Cluster B

Cluster B is composed of 1179 hexagonal regions within the MRSP (Table 2). An intervention immediately at the outset of an epidemic will on average delay infections by 8.17 days for regions in Cluster B. To gain 8 days until the first case, on average, a mobility reduction resembling the eventual response to COVID-19 would need to take place within 3 days of the initial outbreak. To gain at least 5 days, an intervention would be required within 13 days. The slope of a line of best fit is -

0.21, demonstrating the diminishing returns of an intervention as more time passes. The median infection delay curve of Cluster B is presented in Figure 15.

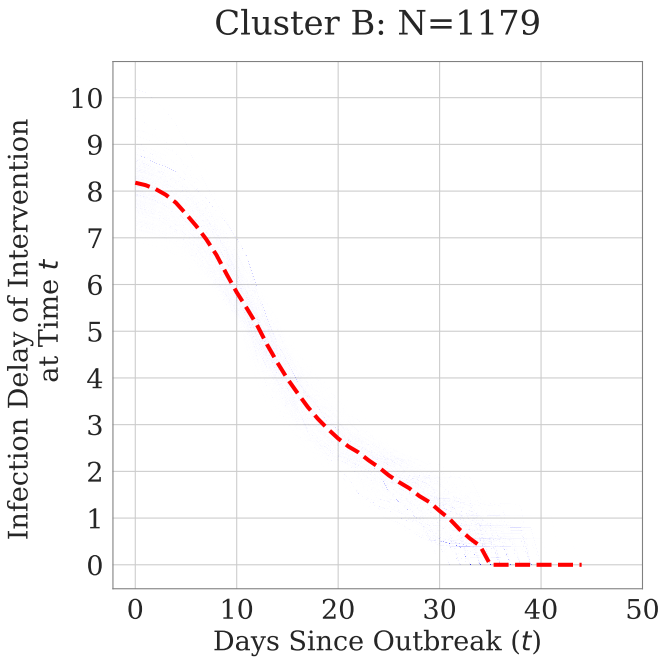


Figure 14: Cluster B infection delay curves (median in red). At time $t = 0$, Cluster B on average gains 8.17 days until its first case following a lockdown.

Cluster B's median in-degree centrality of 567 and out-degree centrality of 566, both at the 47th percentile (Table 3). A median of 3138.16 individuals commute outwards per day, at the 48th percentile, and its median commuter inflow is 13807.57, at the 49th percentile. The median population of regions in Cluster B is 4468.21, at the 53rd percentile. The median income is 1331.55, ranked at the 48th percentile. The black and brown population proportions are 6% and 33%, at the 54th and 52nd percentile, respectively. The white population proportion is 56%, at the 46th percentile. The proportion of individuals with a functioning water network is 98%, at the 52nd percentile. The demographic and centrality percentile characteristics of Cluster B are presented visually in Figure 15. Shown in Figure 11, Cluster B regions are most often found on the outer-middle of the metropolitan area. Relative to Cluster A, they have a greater presence in the near-center area of the MRSP, particularly on the Western half of the region.

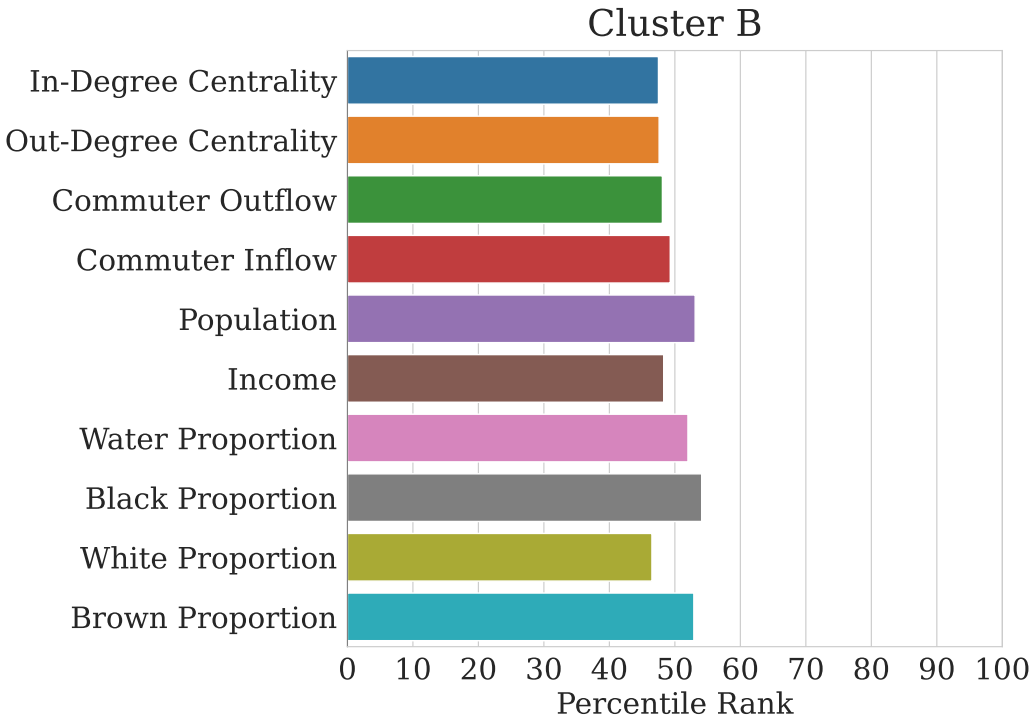


Figure 15: Cluster B: Percentile rank of median centrality and socio-demographic characteristics, relative to all interpolated MRSP regions. All centrality and socio-demographic characteristics fall close to the median.

Cluster C

There are 1008 hexagon regions in the MRSP whose median infection delay curves were assigned to Cluster C (Table 2) For these regions, an intervention at time $t = 0$ can delay infection by 7.08 days on average – the lowest value of all clusters. This means that, on average, mobility-based NPIs are least effective for regions in Cluster C. The median infection delay curve for Cluster C is never above 8 days, meaning that on average, no intervention is able to delay the eventual infection onset by over 8 days. To delay the infection of a disease by at least 5 days, an intervention is required on or before the 13th day from an outbreak. The slope of the line of best fit is -0.18 for the median infection delay curve, meaning that on average, the diminishing returns of delaying an intervention are the least of all clusters. The median infection delay curve of Cluster C is presented in Figure 16.

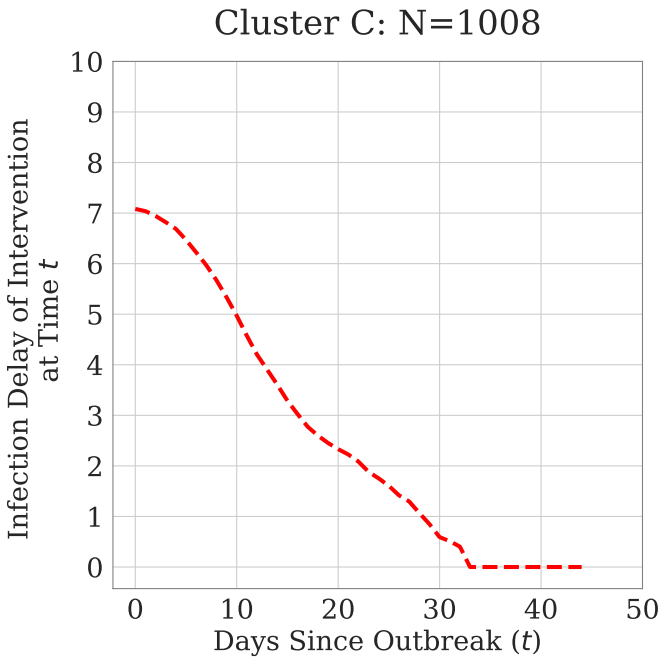


Figure 16: Cluster C infection delay curves (median in red). At time $t = 0$, Cluster C has the lowest median infection delay curve of all clusters, on average gaining 7.08 days until its first case following a lockdown.

Cluster C has the highest median income of 788, at the 77th percentile of all regions in the MRSP (Table 3). Cluster C's median in-degree centrality of 788 is at the 77th percentile, and its out-degree centrality of 796 is at the 78th percentile. Its median commuter outflow is 7360.12, the 66th percentile, and its aggregate commuter inflow has a median of 21343.39, the 70th percentile. The median population of the regions in Cluster C is 4565.34, the highest of all clusters and above the median of the MRSP hexagon regions (54th percentile). The average black, white, and brown percentages are 5%, 64%, and 25%, respectively. The white population percentage is the highest of all clusters at the 63rd percentile across the MRSP. The brown and black proportions are at the 42nd and 35th percentile. The proportion of those with a functioning water network is 0.98, at the 55th percentile rank. The percentile ranks of Cluster C's median demographic characteristics are shown in Figure 17. Figure 11 illustrates that the regions in Cluster C are predominantly focused in the central area of the MRSP. There are also some concentrations in less geographically central regions, such as the collection in the Northeastern corner.

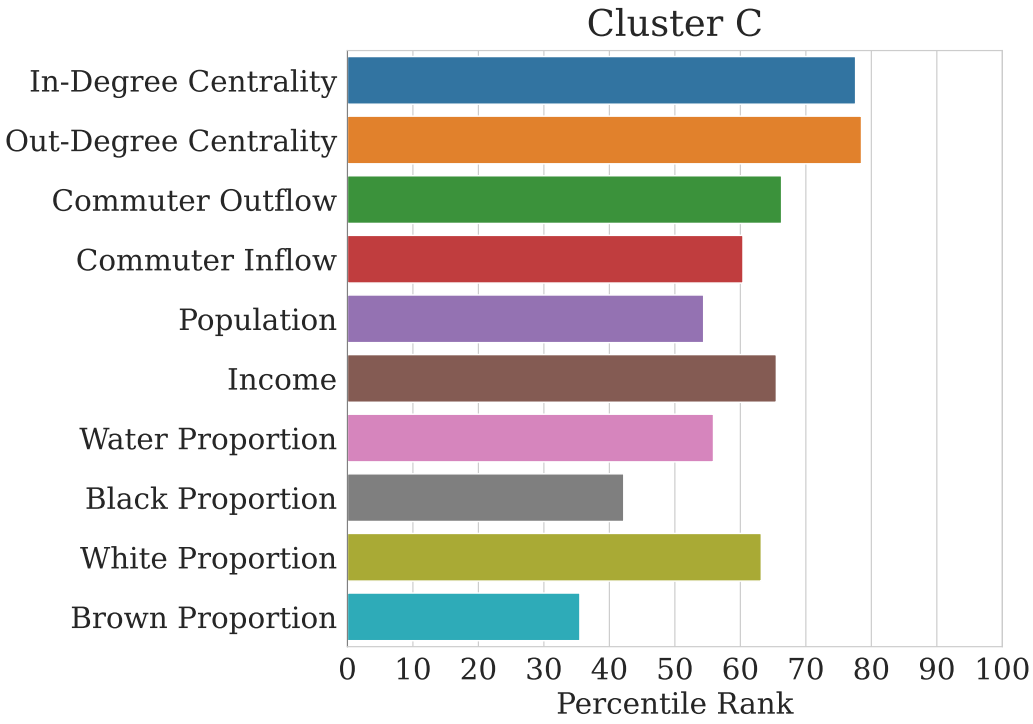


Figure 17: Cluster C: Percentile rank of median centrality and socio-demographic characteristics, relative to all interpolated MRSP regions. In-and-out-degree centrality have the highest relative rankings, and black and brown population proportions have the lowest relative rankings.

3.3 Outlying Outbreak Region Analysis

This section addresses Research Question 2 and highlights the outbreak regions where lockdowns lead to disproportionately high and low infection delay curves across the MRSP. Each outbreak scenario is calculated using the SIR model from Section 2.4. For each region in the MRSP, we identify the outbreak locations leading to infection delay curves above and below the median at time $t = 0$ by one standard deviation. Curves above the median (positive outliers) represent outbreak regions where mobility reductions would lead to greater slowdowns across the MRSP, and vice versa for curves below the median (negative outliers). The p-value on the two-sample t-test of significance is less than 0.01 for every characteristic, and therefore the expected value of positive and negative outliers on each characteristic is statistically significant at the $p < 0.01$ level. Table 4 displays the median characteristics for the positive (+SD) and negative (-SD) outbreak regions, and Figure 18 provides a visual representation of their median percentile ranks.

Median (percentile)		
	Positive Outliers	**Negative Outliers**
In-Degree Centrality***	611.00 (56.15)	583.00 (51.18)
Out-Degree Centrality***	603.00 (54.95)	584.00 (51.44)
Commuters In***	13740.35 (49.14)	10475.37 (42.63)
Commuters Out***	4842.67 (58.05)	2042.05 (40.08)
Population***	3680.56 (48.16)	3190.75 (44.88)
Income (R$/pcm)***	2209.96 (76.97)	1157.07 (35.08)
Water Propor.***	0.96 (52.12)	0.96 (41.37)
Black Propor***	0.03 (17.83)	0.06 (57.82)
White Propor***	0.70 (73.17)	0.51 (33.94)
Brown Propor***	0.19 (24.85)	0.37 (63.53)
Two Sample T-Test Significance:	*p<0.1; **p<0.05; ***p<0.01	

Table 4: Characteristics of positive (blue) and negative (red) outlying outbreak regions. Median and percentile rank relative to all MRSP hexagon regions presented.

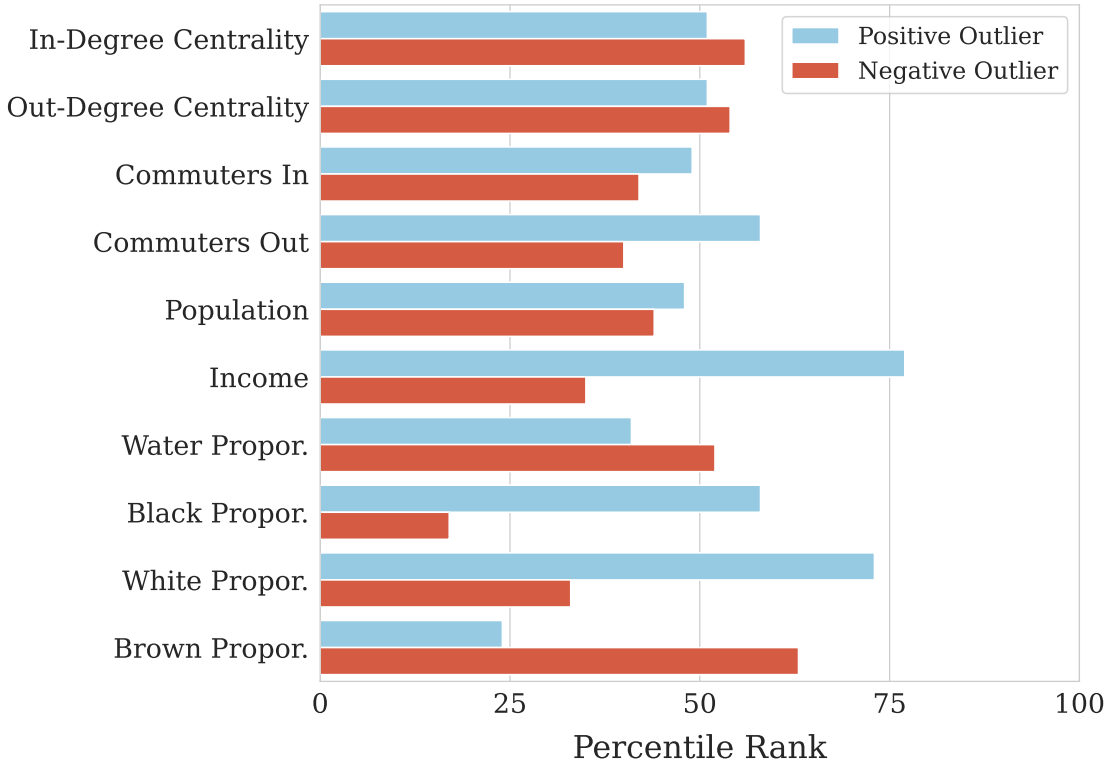


Figure 18: Socio-demographic and centrality percentile ranks comparing positive and negative outliers. Positive outliers, on average, are relatively more white and high-income. Negative outliers, in contrast, have relatively more black and brown individuals, with lower income.

Positive (+SD) outlier outbreak regions lead to greater infection delays across the MRSP when mobility reductions are implemented early-on. They tend to have

a slightly above average centrality in the commuting network, with an average in-degree of 611, and out-degree of 603, at the 56th and 54th percentile, respectively. Observing the volume of commuters, the inward flow is below the median at the 49th percentile, and the commuter outflow is above average at the 58th percentile. On average, these regions have populations below the median, at the 48th percentile. They are also generally of high income, at the 76th percentile. Their median proportion of households with a functioning water network is at the 52nd percentile. The average white population proportion is 0.70, which is at the 73rd percentile. Their black and brown population proportions are 0.03 and 0.19, at the 17th and 24th percentile, respectively.

Negative (-SD) outlier outbreak locations lead to lower infection delays across the MRSP when mobility reductions are implemented early-on. The negative outlier groups tend to have in-and-out-degree centrality near the median, with commuters arriving from 583 distinct regions (in-degree) and travelling to 584 distinct regions (out-degree), both at the 51st percentile. The commuter volumes of 10475.37 inwards and 2042.05 outwards are at the 42nd and 40th percentiles, respectively. Their populations are slightly below the median, at an average of 3190.75. Their median income of 1157.07 at is the 35th percentile, and the median proportion of households with functioning water networks is at the 41st percentile. They have greater black and brown population proportions than those in the positive outliers, at the 57th and 63rd percentiles of all MRSP regions, respectively. They are of low white population proportions, at the 33rd percentile.

Figures 19 and 20 provide a visual representation of the most consistently positive and negative outlier outbreak regions. Each hexagon is assigned a value based on the proportion of total regions that have an above average infection slowdown when it was an outbreak location. For example, if 80% of all regions had an above average (+SD) infection delay curve when region 0 was the outbreak location, region 0's proportion would be 0.8 in Figure 19. This same calculation is performed for Figure 20, except with the below average infection delay curves.

We observe from Figure 19 that outbreaks beginning in central and Western regions more consistently lead to greater infection slowdowns following an early intervention. When these regions are outbreak locations, the relative effectiveness of an early mobility-based intervention is increased for many regions across the MRSP. Based on the demographic variable analysis provided in Table 4, these regions are more predominantly white and high income.

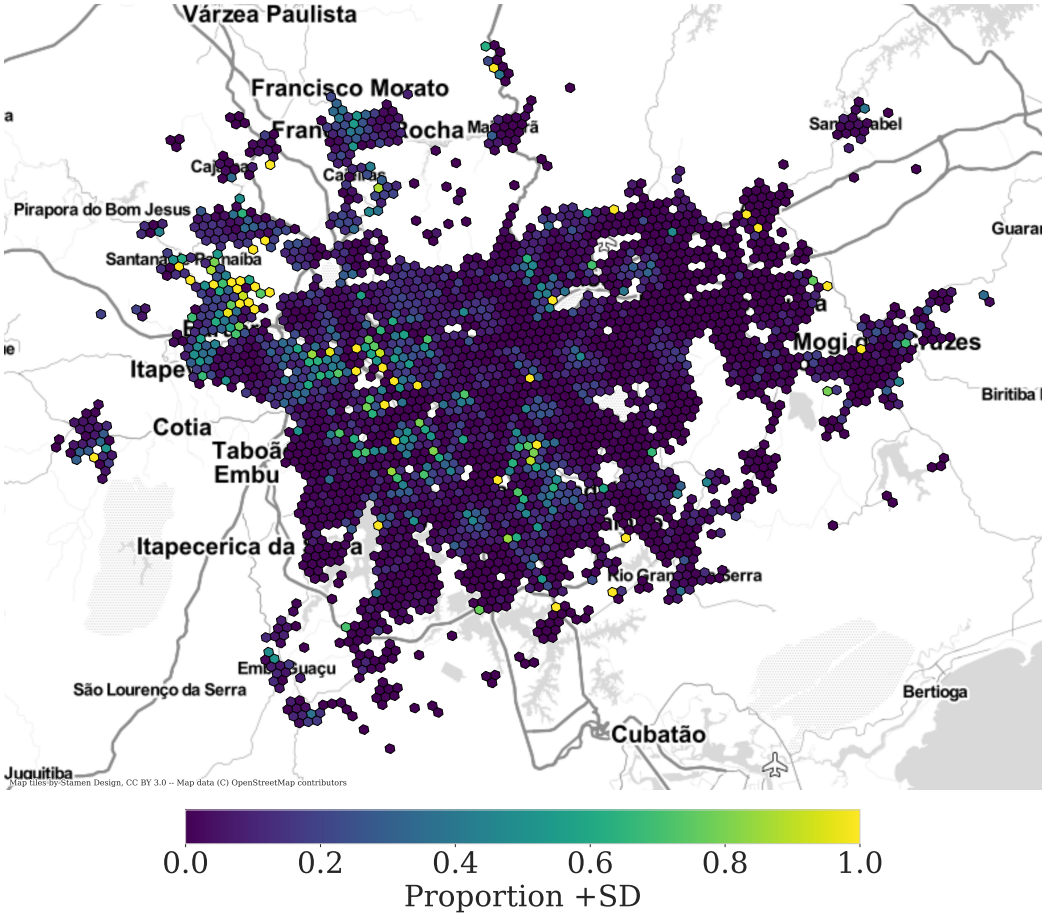


Figure 19: Positive outlier display. Colours assigned according to proportion of regions with *above* average (+SD) infection delay when that region is the outbreak location. Outbreaks beginning in lighter hexagons can be slowed to a *greater* extent across the MRSP following a lockdown.

Figure 20 depicts an equivalent plot for the negative outliers, representing the outbreak locations where interventions were relatively less effective. An outbreak location's value is assigned based on the proportion of all regions that gain less time to a first case from an intervention, relative to all 2737 outbreak scenarios. Relative to the positive outliers, outbreak locations that consistently lead to lower infection delay curves are more concentrated in the Western part of the MRSP, with minor presence towards the center. Based on the results of Table 4, these regions are predominantly of greater black and brown population proportions and lower income.

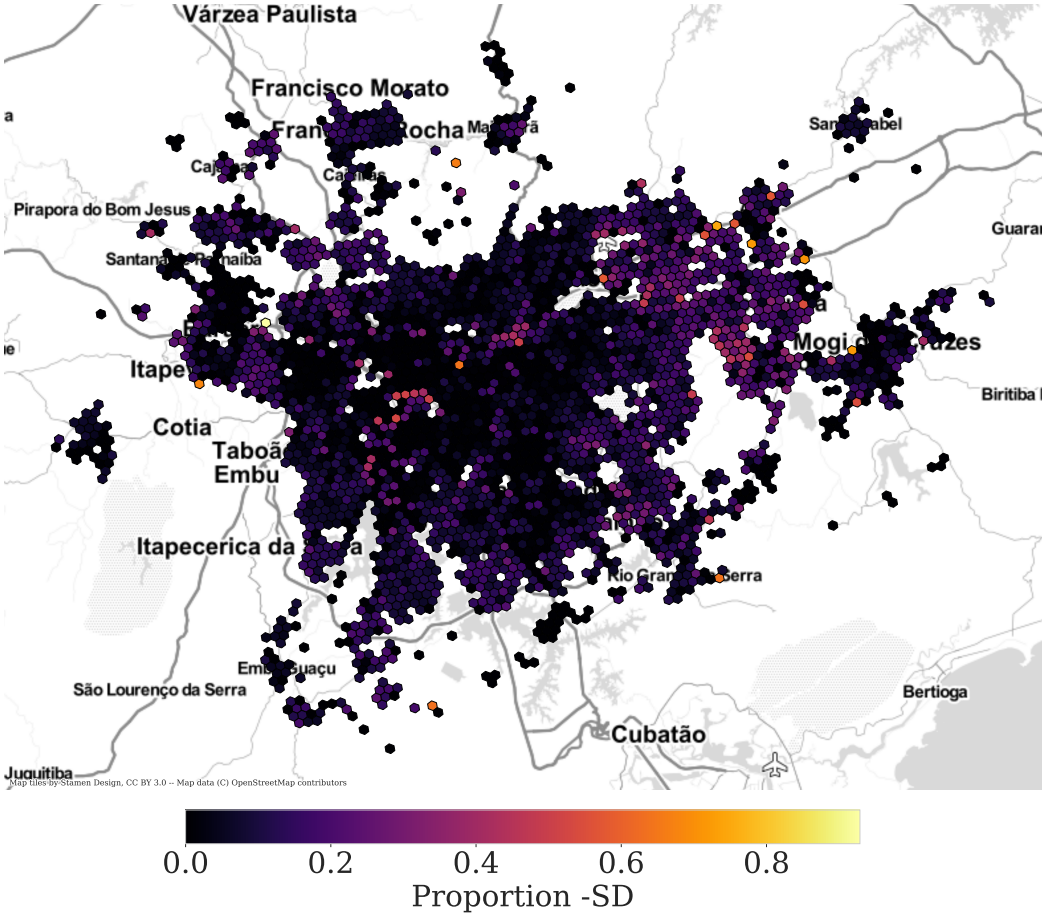


Figure 20: Negative outlier display. Colours assigned according to proportion of regions with a *below* median (-SD) infection delay when that region is the outbreak location. Outbreaks beginning in lighter hexagons can be slowed to a *lesser* extent across the MRSP following a lockdown.

4 Discussion

In this paper, the novel Infection Delay Model has been developed to understand lockdown effectiveness in delaying a region's first case using commuting and cellphone mobility data. With effective distance-based metrics to calculate arrival times and infection delays, the model presents an intuitive understanding of lockdown effectiveness, visualised with infection delay curves. This approach, applied to simulated epidemics across the MRSP, contributes to a forward-looking understanding of disease risk and regional benefits to lockdowns. Integrating cellphone mobility data mimics regional changes made in the MRSP before and after their initial COVID-19 lockdown on March 16, 2020. By contextualising the model's results with census

data, new insights into the socio-demographic patterns of lockdown effectiveness have been found.

After validating the relationship between effective distance and arrival times, the ability of lockdowns to delay disease spread in the MRSP is analysed from two points of view: regions-at-risk and regions of outbreak. The first analysis (k-means) observes the interpolated hexagons in the MRSP as susceptible regions not yet infected, and assesses how much a first infection can be delayed by interventions over the course of simulated outbreaks. The second analysis (outlying outbreak regions) looks at the epidemics simulated from each region, and assesses the degree that spread can be slowed to the rest of the MRSP from an early lockdown intervention.

4.1 Effective Distance and Arrival Time Validation

Validating the relationship between effective distance metrics and disease arrival times is necessary to justify its use in the IDM. Using an estimated *first* COVID-19 arrival date and its specific location from the São Paulo SARI database, this process assessed the relationship between the 2737 regions' effective distance from this outbreak location, and its eventual time to arrival. The analysis confirmed that larger effective distances from the outbreak location lead to later arrival times in the social isolation commuting network. This matches the findings of Brockmann and Helbing (2014) and Iannelli et al. (2017), regarding the predictive ability of effective distance metrics for disease arrival insights. The positive relationship between effective distance and COVID-19 arrival times was statistically significant in the MRSP with or without an additional control for regional populations. This validation is crucial to the subsequent use of effective distances to compute the arrival times and infection delay curves. Additionally, the adjusted radiation model applied to estimate 2020 commuting flow, and the aerial weighting interpolation to the hexagons, was sufficient to predict commuting-based disease dynamics at the hexagon-level.

4.2 Clusters and Intervention Effectiveness

To create a generalised model of early intervention effectiveness, this analysis takes into account 2737 simulated outbreak locations within the MRSP, with regions clustered according to their overall benefit from lockdowns interventions. Further, the mobility data from the COVID-19 pandemic, applied to the intervention scenarios, allows for neighbourhood-level heterogeneity in a pandemic context. Using the IDM to generate effective distance-based infection delay curves, three distinct socio-demographic groups were identified.

- ***Cluster A*** – *Low Centrality, High Socio-Demographic Vulnerability*: This cluster consists primarily of regions on the outer edges of the relevant MRSP area, with very low centrality and low commuting inflow and outflow. It has the highest black and brown population proportions of all regions, with comparatively lower population, income, functioning water networks, and white population.

- ***Cluster B*** – *Average Centrality, Average Socio-Demographic Vulnerability*: This cluster has high presence on the periphery of the MRSP, with penetration towards the outer-middle area. Its centrality and socio-demographic characteristics percentile ranks are close the MRSP medians. It has slightly above average populations, black and brown population proportions, and functioning water network proportions. It has slightly below average centrality, commuter inflow and outflow, white proportion, and income.

- ***Cluster C*** – *High Centrality, Low Socio-Demographic Vulnerability*: This cluster has high presence in the central MRSP, with minor presence towards the periphery. It is characterised by high centrality and high commuter inflow and outflow. It also has the highest white population proportions, overall populations, income, and functioning water network proportions. Cluster C has the lowest brown and black population proportions of all regions.

Assessment of the infection delay curves for these three clusters reveals that early interventions delay infection to a greater extent in less central, and more socio-demographically vulnerable regions. Socio-demographic vulnerability is reflected by regions with lower income, lower functioning water networks, and greater black and brown population proportions. This finding supports H_1 that infection delays will be most prominent in less central, and more socio-demographically vulnerable regions.

Cluster A has the lowest income, proportion of individuals with functioning water networks, and the largest black and brown population of all clusters. Observing the commuting network, they have the lowest centrality and aggregate commuter activity. The income and racial information of this cluster most closely resemble the socially vulnerable groups discussed by Rocha et al. (2021) regarding the impacts of inequality on health outcomes during the 2020 epidemic in the MRSP. The fewer hospital resources and physicians in disadvantaged regions were found to lead to more rapid case acceleration after the initial outbreak of COVID-19 (Rocha et al., 2021). The unequal acceleration of cases in such regions emphasises the importance of mobility-based interventions to delay the arrival of diseases and improve preparatory measures. Disproportionately low income, black, and brown regions

were found to have greater hospitalisation and death rates from COVID-19 in Li, Sabrina, Pereira et al. (2020), and importantly appear to benefit the *most* from early mobility-based intervention based on this analysis.

Cluster C is the most populated, central, wealthy, and white of all k-means groups. In the early stages of the COVID-19 pandemic, wealthier and whiter regions faced higher hospitalisation rates due to international introductions from higher-income travellers (Li, Sabrina, Pereira et al., 2020). The interconnectedness of the densely populated and high-income regions makes them especially vulnerable to infection during the outbreak of disease. Assessing how these regions would react under varying outbreak scenarios is useful for broadening our understanding of disease resilience. Given that diseases or variants may emerge locally or from non-major transport routes, these socio-demographically secure and highly connected regions benefit relatively less from a mobility-based intervention of the scale observed during the COVID-19 epidemic.

Cluster B's in-and-out-degree centrality, commuting inflow and outflow, and socio-demographic characteristics are close to the MRSP's medians, lying between the extremes noted in clusters A and C. Less vulnerable than Cluster A, it has higher white population proportions relative to black and brown groups, and a higher median income. Observing its median infection delay curves, lockdowns delay the infection of Cluster B regions less than Cluster A, but more than Cluster C. The association between its vulnerability and lockdown effectiveness follow the hypothesised trend, placing it between the two clusters on both measures.

The IDM's infection delay curves are based on the commuting network of the MRSP and the mobility reductions of infected regions over the course of simulated outbreaks. As such, this analysis is not making causal claims regarding socio-demographic vulnerability and early intervention effectiveness. Using a holistic lens that takes into account the commuting network, socio-demographic characteristics, and social isolation capabilities of regions, this thesis characterises the varied impact of early interventions across the MRSP. The value in this analysis is its demonstration that even when looking across every possible outbreak scenario in the MRSP, there are identifiable differences in how much regions benefit from early interventions, and that these differences intersect with real socio-demographic inequalities. The greater effectiveness of earlier lockdowns in socially vulnerable clusters can help inform policy-makers seeking to minimise the health burden across the MRSP.

These findings favour swift mobility reductions immediately after an initial outbreak in the MRSP. After a lockdown, socio-demographically disadvantaged groups tend to have greater infection delays relative to other regions, a critical finding to

equalise the health inequalities in Brazil – a growing priority for Brazilian policy-makers (Landmann-Szwarcwald & Macinko, 2016). Additionally, the infection delay curves in more vulnerable clusters have steeper median slopes (Table 2), meaning they tend to lose more every day a lockdown is not implemented. The MRSP-specific findings on the effectiveness of interventions can be translated into highly-targeted policies, as each Brazilian state operates like its own country in a federalised system (Neiva et al., 2020). For these recommendations to be economically feasible, economic safety nets must be properly targeted to support the vulnerable communities most burdened by the lockdown (Lee et al., 2020). Another important factor in the effectiveness of such mobility-based interventions is the consistency of messaging. Jair Bolsonaro, who was deemed by the *Lancet* editors as the greatest threat to Brazil's COVID-19 response, actively disregarded such policy measures, sending conflicting messages across the country as state authorities implemented lockdown measures (The Lancet, 2020). In addition to early lockdown measures, coherent and targeted messaging is essential to maximise effectiveness.

4.3 Outlying Outbreak Regions and Socio-Demographics

Taking the perspective of outbreak regions, this second analysis determines the features of outbreak locations where lockdowns can most and least slow the spread of disease. For every region at risk, the outbreak locations where lockdowns lead to higher and lower infection delay curves are identified and their characteristics aggregated. The differences between the positive and negative outlier outbreak regions are most evident on their socio-demographic characteristics, a finding made possible through the inclusion of cellphone mobility data.

The greatest slowdowns across the MRSP, identified by the positive outliers, occurred in the regions of higher affluence. In line with H_2, these occurred in areas of higher income, functioning water networks, and white population proportions, with lower black and brown proportions. These results are supported by the findings of Li, Sabrina, Pereira et al. (2020), who show that affluent regions in the MRSP have higher social isolation capacities. When affluent regions are the source of an outbreak, their greater mobility reductions translate to more days between baseline and intervention arrival times for the rest of the MRSP. In the effective distance-based arrival time calculations, the mobility reductions of an outbreak location dominate how a lockdown can slow down the spread, overriding factors related to their centrality in the network.

In contrast, the least effective slowdowns across the MRSP, identified by the negative outliers, occurred in regions of higher vulnerability, further supporting

H_2. The negative outlier outbreak regions are above average in black and brown populations, and well below average in income and white populations. The findings of this thesis expand on Li, Sabrina, Pereira et al.'s (2020) findings that vulnerable regions are less able to self-isolate after a lockdown, showing that these inequalities in social-isolation capabilities, captured by cellphone mobility data, manifest into concrete differences in early intervention effectiveness.

To illustrate the importance of using cellphone mobility data in the IDM, the negative outliers when *applying a uniform mobility reduction* are displayed in Figure 21. Using this simplistic, blanket mobility reduction, higher centrality emerges as a dominant factor in outbreak locations where lockdowns have less effectiveness. The true analysis, presented in Figure 20 and Table 4 *with cellphone data*, captures the socio-demographic dimension of mobility, and paints a more complete picture. It in fact shows centrality and commuting flow *below* the median in negative outliers, and high income and white population proportions become the largest differentiators.

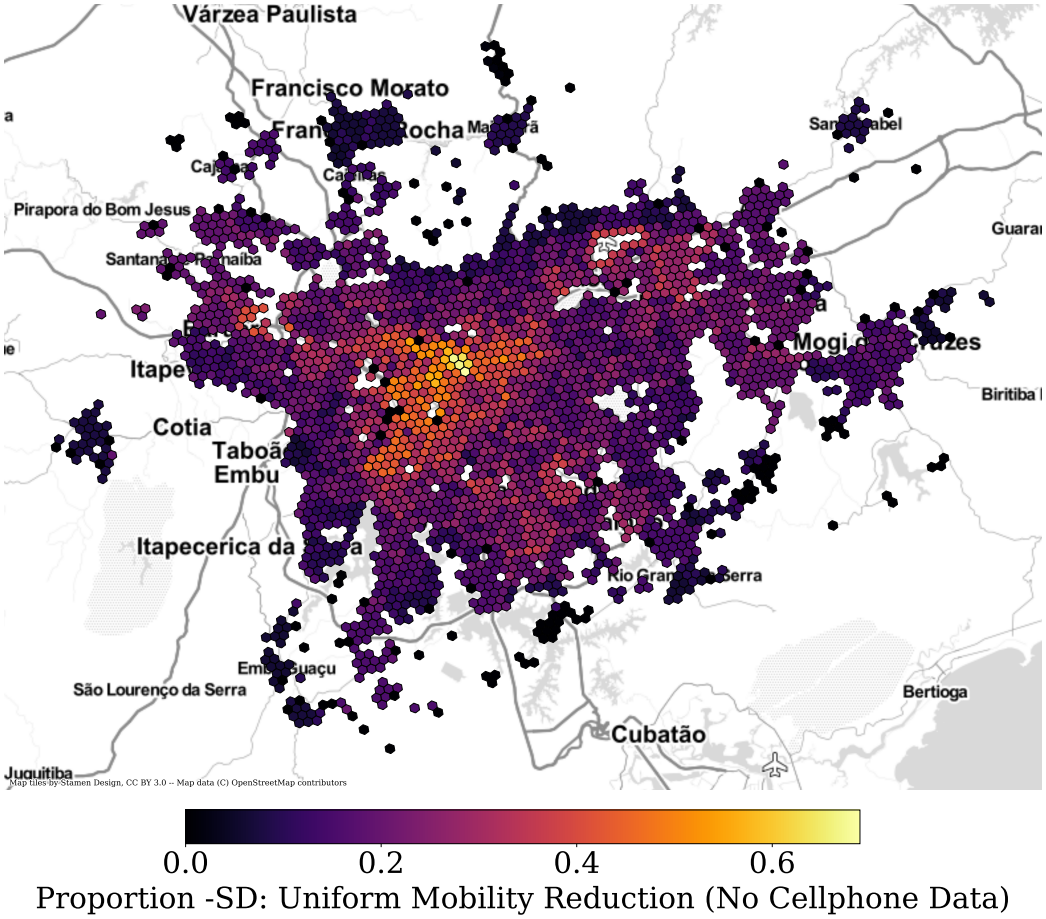


Figure 21: Negative outlier display for uniform mobility reduction by 55% (mean mobility reduction across all regions). Colours assigned according to the proportion of regions with an *below* median (-SD) infection delay when that region is the outbreak location. Outbreaks beginning in lighter hexagons can be slowed to a *lesser* extent across the MRSP following a lockdown. Centrality of outbreak locations in the MRSP drives lockdown ineffectiveness.

According to this work, intervening early when a disease outbreak begins in positive outlier regions yields a greater return on investment for delaying the spread to other regions in the MRSP. While this seems to indicate that more emphasis should be placed on an early intervention when outbreaks begin in the positive outlier locations, it is important that the findings are contextualised with social data. Outbreak locations that tend to lead to higher infection delay potential for the rest of the MRSP are disproportionately wealthy and white regions in the state. This indicates that the regions that tend to face the lowest health burdens during an outbreak, according to Li, Sabrina, Pereira et al. (2020), are the outbreak locations

commonly associated with above average intervention effectiveness for the rest of the MRSP.

The results of the outlying outbreak region analysis highlight that *where* an epidemic begins, even on an intrastate level, can convey information about how much a disease can be slowed down by an early mobility-based intervention. The findings of this section show that inequalities in mobility reductions impact the ability of interventions to slow down state-wide spread, and that governments should focus on providing a safety net to allow more vulnerable regions to self-isolate without sacrificing their livelihood. The systematic variability in intervention effectiveness based on the outbreak location is found to intersect with key socio-demographic characteristics, and demonstrates the value of simulating varieties of pandemic scenarios when evaluating lockdown effectiveness.

4.4 Limitations

Cellphone Data Coverage Bias

When considering cellphone data sources, originally collected for commercial purposes, coverage bias should be noted. As a cellphone analytics company, the sample of users in the In Loco data set is determined by their market share, rather than an emphasis on representative samples (Tizzoni et al., 2014). The near-global ubiquity of cellphones does not preclude biases, as possession and use rates vary across demographic and income groups (Kraemer et al., 2020). The elderly are often underrepresented in such samples, while educated urban males are overrepresented relative to lower-income individuals (Kraemer et al., 2020). As In Loco collects app-based cellphone location data, income-related biases are further exacerbated, as the partnered apps may require smartphone functionalities or access to a wireless network to connect to the internet (Oliver et al., 2020). The movements of these users beyond their residential areas, required to calculate the social isolation values, may not be representative of the region as a whole, due to the aforementioned demographic and income biases.

The social isolation regions used in this analysis are heavily concentrated towards the municipality of São Paulo and its surrounding regions, with more sparse presence of hexagons on the periphery of the MRSP (Figure 3). This immediately excludes large portions of non-central MRSP from the infection delay analysis based on real mobility changes. Although the primary urban regions are captured by the data set, the undercoverage of the cellphone mobility data limits the scope of the analysis.

Social Isolation and Commuting Mobility

After calculating the marginal change in social isolation during the pre-and-post-intervention periods in the MRSP, we take this *mobility* reduction as a proxy for general *commuting* reduction in Equation 7. It is assumed that the proportional commuting flow remains constant, and that there are simply less individuals in circulation. In reality, as the pandemic changed commuting more than scaling down the number of commuters. This may introduce a specification error, defined by Amaya et al. (2020) as the failure of the measurement (social isolation) to capture the required concept (commuting changes). As an example, teleworking access is a key factor in shaping the daily commuting patterns of the MRSP residents, as some occupations require in-person services while others can be conducted remotely (Li, Sabrina, Pereira et al., 2020). The distribution of occupations which can transition to teleworking is not evenly distributed across the MRSP, such as informal occupations, concentrated in lower-income regions (Moreno-Monroy & Ramos, 2020).

The suitability of integrating traditional commuting data with the aggregated social isolation cellphone data deserves exploration by future research. This study recommends comparing granular cellphone mobility data, such as the location pairs used in Peixoto et al. (2020), and observing how daily commuting patterns and their changes after lockdown resemble those found in this paper's analysis. If providing similar results, the greater anonymity of the aggregated social isolation data may render it a more readily accessible and minimally invasive method for granular mobility-related studies.

Homogeneity of Demographic Characteristics

Interpolating the demographic and commuting flow data to the social isolation hexagons was an integral component of this analysis. It allowed the infection delay analysis to involve real-world mobility changes when calculating the effective-distance based arrival times under an intervention scenario. The aerial weighting method used in this analysis is highly transparent, yet relies on a strong assumption of homogeneously distributed demographic characteristics. When using the degree of overlap with one set of regions to estimate values for another set of regions, it is implicitly assumed that the values in the known regions are evenly distributed across their area (Huyen Do, Thomas-Agnan, & Vanhems, 2015). For example, if a social isolation hexagon covers 5% of a region that has a population of 100, the hexagon is assigned a population of 5. In reality, the distribution of the population may be heterogeneous, with more or less than 5 individuals in the area overlapping with the hexagon. Homogeneity is a rare occurrence in the real world, and while

dominant trends are captured by the interpolation method, there will be regions inaccurately allocated demographic characteristics (Lam, 1983).

Using Cluster Medians in Infection Delay Analysis

To develop the three infection delay clusters, the median curves of each region were calculated (shown in Figure 3), and then subsequently clustered. The median of the cluster was then taken to develop summary statistics and characterise its features. The heterogeneity within each cluster is reduced to a single infection delay curve, useful for characterization but failing to account for how curves within a cluster may differ. The shaded blue lines in Figures 12, 14, and 16 show that substantial variation exists within each cluster, straying from the median. This paper uses the outbreak outlier analysis to understand the characteristics of outbreak locations that lead to within-region variation, but a more general method unsupervised learning method could be developed. An analysis was attempted clustering all 2737×2737 infection delay curves, but a k-means analysis was unable to detect meaningful intervention effectiveness or social characteristic differences across the 7.4 million curves. This is most likely attributed to the raw volume of observations in the clusters obscuring differences in demographic characteristics.

Transport Types

Mode of transport is a highly influential feature of commuting-based disease spread. Whether one drives a private car, bikes, or takes public transport has implications for their exposures on commuting routes (Francetic & Munford, 2021). The effective distance calculations used to generate arrival times and the SIR models used to simulate pandemic scenarios are unable to differentiate between transport types. Due to this limitation, aggregate flows of all transport types were used to construct the commuting network, failing to include the mode of transport. Future studies could separate the commuting data by mode of transport and see how predictive their effective distance metrics are for arrival times, and use this to develop a transport-type specific model of disease spread. The movement of individuals across the MRSP is accurately captured by the commuting analysis, yet a further emphasis on *how* they move would add more depth to a commuter-oriented epidemiological analysis.

4.5 Conclusion

Urgent research into the effectiveness of government interventions to slow disease spread is essential, as the humanitarian catastrophe resulting from the COVID-

19 pandemic and its emerging new variants continues globally. The Infection Delay Model, as part of a generalised method of assessing intervention effectiveness, focuses on the ability of mobility reductions to slow the spread of infection. The data-linkage approach, interpolating commuting and socio-demographic data, allowed for insights into the social context of regions and how interventions can delay eventual infections. The unique integration of cellphone mobility data into the effective distance metrics has captured heterogeneous changes in isolation, found in prior literature to intersect with socio-demographic inequalities (Li, Sabrina, Pereira et al., 2020; Lee et al., 2020). While this analysis is focused on Brazil, a region where income, health, and transport inequalities are stark, as discussed in Malta et al. (2020), the presented approach can be applied in other regions to observe the intersection of intervention effectiveness and socio-demographic vulnerability. Similarly, the epidemiological parameters in this analysis are chosen to mimic COVID-19, but a novel variant or disease's reproduction rate and infectious period could be used as substitutes.

The infection delay curves for each cluster efficiently and interpretably summarise regional benefits of early interventions across 2737 outbreak scenarios, and their intersections with socio-demographic information. The regions gaining the most from early interventions are socio-demographically vulnerable regions of low centrality, conveying that early interventions can contribute to equalising the outsized health burden of those less prepared to weather an epidemic. From an outbreak location perspective, lockdowns can delay the spread of disease to a greater extent when more affluent and white regions are the first sites of a disease. This is consistent with prior literature on the greater ability of wealthy and whiter areas to self-isolate, which this analysis captures by including of cellphone mobility data (Li, Sabrina, Pereira et al., 2020).

The Infection Delay Model, a major contribution of this study and a novel approach, provides a comprehensive method of assessing early intervention effectiveness. It is known from prior literature that social isolation levels vary based on socio-economic status, and that these differences impact disease exposure and health burdens (Li, Sabrina, Pereira et al., 2020; Lee et al., 2020). This method builds on prior literature and, by merging social isolation and traditional data sources, determines how these inequalities tangibly manifest into early intervention effectiveness for future epidemics. The backbones of the analysis are publicly available census and commuting surveys, which can be scaled to the relevant year, in addition to social isolation data, which is aggregated and highly anonymised. Using a combination of these three secondary data sources, a holistic picture of the MRSP is developed, providing a nuanced perspective of the infection delays caused by early

interventions. The replicable methods generated in this study are relevant for global decision-makers seeking to generate socially relevant insights on intervention decisions.

A second major contribution of this study is its generalised, forward-looking characterisation of intervention effectiveness. Rather than relying on a single set of initial conditions when modeling a disease, or dividing the paper between a subset of transport hubs, this analysis incorporates *all* possible outbreak locations when assessing how early interventions lead to infection delays. This allows for broad understandings of intervention effectiveness whose validity is not reliant on the next epidemic beginning in the same location as the last. This generalised approach makes the outlying outbreak region analysis possible, where regions of outbreak are assessed based on the extent to which early lockdowns led to greater and lower infection delays for the rest of the MRSP. This is important from a preparedness standpoint, allowing policy-makers to understand the relative gains to be made from lockdowns based on the outbreak location, combined with a system of rapid testing, mobile vaccination units, and information campaigns. From a social equity standpoint, it illustrates how mobility reduction inequalities translate into lower lockdown effectiveness for the rest of the MRSP, placing all regions at greater risk when outbreaks begin in regions with lower abilities to reduce mobility.

Taking into account all outbreak locations, this study provides evidence of fundamental relationships between socio-demographic inequalities and intervention effectiveness – a phenomenon for future research to explore in other global contexts. The unequal global burden of COVID-19 has emphasised the importance of factoring social data into health policy and preparedness decisions, dramatically impacting individuals' lives. Adopting interdisciplinary methodologies to investigate the effectiveness of interventions, with a focus on exploring inequalities, may provide novel insights into the factors driving the unequal playing field exposed during the COVID-19 pandemic. Inequalities cannot be addressed until they are recognised, which this study's data-driven but socially-grounded approach has sought to uncover.

Appendices

A Hexagon-Interpolated Summary Statistics

	In-Degree	Out-Degree	Commute In	Commute Out
Mean	617.59	617.41	23385.11	23385.11
Std.	278.40	279.77	29771.83	101752.19
Min.	57.00	57.00	6.51	18.01
25%	436.00	431.00	3646.88	685.03
50%	580.00	580.00	14156.38	3416.13
75%	758.00	757.00	32724.17	10627.40
Max.	2228.00	2101.00	260048.72	2506462.05
	Population	Income	Water Propor.	
Mean	4549.95	2116.47	0.90	
Std.	3288.77	2250.19	0.177	
Min.	119.83	0.00	0.00	
25%	1672.34	1061.98	0.901	
50%	4373.74	1386.62	0.98	
75%	6783.39	2129.54	0.99	
Max.	21351.28	28471.98	1.0	
	Black Propor.	White Propor.	Brown Propor.	
Mean	0.05	0.59	0.3	
Std.	0.02	0.15	0.13	
Min.	0.00	0.00	0.01	
25%	0.03	0.47	0.18	
50%	0.05	0.58	0.31	
75%	0.07	0.71	0.41	
Max.	0.41	0.95	0.65	

Table 5: Summary statistics of hexagon-interpolated socio-demographic traits

References

Aleta, A., Mart, D., Bakker, M. A., Pastore, A., & Ajelli, M. (2020). Quantifying the importance and location of SARS-CoV-2 transmission events in large metropolitan areas. , 1–18.

Amaya, A., Biemer, P. P., & Kinyon, D. (2020). Total Error in a Big Data World: Adapting the TSE Framework to Big Data. *Journal of Survey Statistics and Methodology*, *8*(1), 89–119. Retrieved from https://doi.org/10.1093/jssam/smz056 doi: 10.1093/jssam/smz056

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., ... Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, *734*, 1–74. Retrieved from https://doi.org/10.1016/j.physrep.2018.01.001 doi: 10.1016/j.physrep.2018.01.001

Brockmann, D., & Helbing, D. (2014). The hidden geometry of complex, network-driven contagion phenomena (Science (1337)). *Science*, *343*(6172), 730. doi: 10.1126/science.343.6172.730-c

Burns, E., Martins, L., Schneider, R. M., Momsen, R. P., & James, P. E. (2021). Brazil. *Encyclopedia Britannica*. Retrieved from https://www.britannica.com/place/Brazil.

Candido, D. S., Claro, I. M., de Jesus, J. G., Souza, W. M., Moreira, F. R., Dellicour, S., ... Faria, N. R. (2020). Evolution and epidemic spread of SARS-CoV-2 in Brazil. *Science*, *369*(6508), 1255–1260. doi: 10.1126/SCIENCE.ABD2161

Chakraborty, C., Bhattacharya, M., Sharma, A. R., Lee, S.-S., & Agoramoorthy, G. (2021, jun). SARS-CoV-2 Brazil variants in Latin America: More serious research urgently needed on public health and vaccine protection. *Annals of medicine and surgery (2012)*, *66*, 102428. Retrieved from https://pubmed.ncbi.nlm.nih.gov/34109031https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8178066/ doi: 10.1016/j.amsu.2021.102428

Chang, S., Pierson, E., Koh, P. W., Gerardin, J., Redbird, B., Grusky, D., & Leskovec, J. (2020). Mobility network models of COVID-19 explain inequities and inform reopening. *Nature*, *589*(January). Retrieved from http://dx.doi.org/10.1038/s41586-020-2923-3 doi: 10.1038/s41586-020-2923-3

Christley, R. M., Pinchbeck, G. L., Bowers, R. G., Clancy, D., French, N. P., Bennett, R., & Turner, J. (2005). Infection in Social Networks: Using Network Analysis to Identify High-Risk Individuals. *American Journal of Epidemiology*, *162*(10), 1024–1031. Retrieved from https://doi.org/10.1093/aje/kwi308 doi: 10.1093/aje/kwi308

Coelho, F. C., Lana, R. M., Cruz, O. G., Villela, D. A., Bastos, L. S., Pastore Y Piontti, A., ... Gomes, M. F. (2020). Assessing the spread of COVID-19 in Brazil: Mobility, morbidity and social vulnerability. *PLoS ONE*, *15*(9 September), 1–11. doi: 10.1371/journal.pone.0238214

Comber, A., & Zeng, W. (2019). Spatial interpolation using areal features: A review of methods and opportunities using new forms of data with coded illustrations. *Geography Compass*, *13*(10), e12465. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/gec3.12465 doi: https://doi.org/10.1111/gec3.12465

Covid: Brazil has more than 4,000 deaths in 24 hours for first time. (2021, Apr). *BBC*.

Coy, M., Sandholz, S., Töpfer, T., & Zirkl, F. (2018, dec). Metropolitan Regions: Challenges for Sustainability and Governance of Periurban Areas in Brazil. doi: 10.4000/confins.16904

de Andrade, C. L. T., Pereira, C. C. d. A., Martins, M., Lima, S. M. L., & Portela, M. C. (2020, dec). COVID-19 hospitalizations in Brazil's Unified Health System (SUS). *PLOS ONE*, *15*(12), e0243126. Retrieved from `https://doi.org/10.1371/journal.pone.0243126`

de Souza, F. S. H., Hojo-Souza, N. S., Batista, B. D. d. O., da Silva, C. M., & Guidoni, D. L. (2021, mar). On the analysis of mortality risk factors for hospitalized COVID-19 patients: A data-driven study using the major Brazilian database. *PLOS ONE*, *16*(3), e0248580. Retrieved from `https://doi.org/10.1371/journal.pone.0248580`

de Souza, W. M., Buss, L. F., Candido, D. d. S., Carrera, J.-P., Li, S., Zarebski, A. E., ... Faria, N. R. (2020). Epidemiological and clinical characteristics of the COVID-19 epidemic in Brazil. *Nature Human Behaviour*, *4*(8), 856–865. Retrieved from `https://doi.org/10.1038/s41562-020-0928-4` doi: 10.1038/s41562-020-0928-4

Dong, E., Du, H., & Gardner, L. (2020, may). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet Infectious Diseases*, *20*(5), 533–534. Retrieved from `https://doi.org/10.1016/S1473-3099(20)30120-1` doi: 10.1016/S1473-3099(20)30120-1

Francetic, I., & Munford, L. (2021). Corona and coffee on your commute: a spatial analysis of COVID-19 mortality and commuting flows in England in 2020. *European Journal of Public Health*. Retrieved from `https://doi.org/10.1093/eurpub/ckab072` doi: 10.1093/eurpub/ckab072

Gautreau, A., Barrat, A., & Barthélemy, M. (2008). Global disease spread: Statistics and estimation of arrival times. *Journal of Theoretical Biology*, *251*(3), 509–522. doi: 10.1016/j.jtbi.2007.12.001

Grantz, K. H., Meredith, H. R., Cummings, D. A., Metcalf, C. J. E., Grenfell, B. T., Giles, J. R., ... Wesolowski, A. (2020). The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nature Communications*, *11*(1), 1–8. Retrieved from `http://dx.doi.org/10.1038/s41467-020-18190-5` doi: 10.1038/s41467-020-18190-5

Guo, Q., Lu, X., Gao, Y., Zhang, J., Yan, B., Su, D., ... Wang, G. (2017). Cluster analysis: a new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients. *Scientific Reports*, *7*(1), 43965. Retrieved from `https://doi.org/10.1038/srep43965` doi: 10.1038/srep43965

Haug, N., Geyrhofer, L., Londei, A., Dervic, E., Desvars-Larrive, A., Loreto, V., ... Klimek, P. (2020). Ranking the effectiveness of worldwide COVID-19 government interventions. *Nature Human Behaviour*, *4*(12), 1303–1312. Retrieved from `https://doi.org/10.1038/s41562-020-01009-0` doi: 10.1038/s41562-020-01009-0

Henry, R. (2020). Covid-19 in latin america: a humanitarian crisis. *Lancet*,

396(10261), 1463.

Hone, T., Stokes, J., Trajman, A., Saraceni, V., Coeli, C. M., Rasella, D., ... Millett, C. (2021). Racial and socioeconomic disparities in multimorbidity and associated healthcare utilisation and outcomes in Brazil: a cross-sectional analysis of three million individuals. *BMC Public Health*, *21*(1), 1287. Retrieved from https://doi.org/10.1186/s12889-021-11328-0 doi: 10.1186/s12889-021-11328-0

Hunter, E., Namee, B. M., & Kelleher, J. D. (2020, apr). A Model for the Spread of Infectious Diseases in a Region. *International journal of environmental research and public health*, *17*(9), 3119. Retrieved from https://pubmed.ncbi.nlm.nih.gov/32365775https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7246727/ doi: 10.3390/ijerph17093119

Huyen Do, V., Thomas-Agnan, C., & Vanhems, A. (2015). Spatial reallocation of areal data – another look at basic methods. *Revue d'Économie Régionale Urbaine*, *mai*(1-2), 27–58. Retrieved from https://www.cairn.info/load_pdf.php?ID_ARTICLE=RERU_151_0027https://www.cairn.info/revue-d-economie-regionale-et-urbaine-2015-1-2-page-27.htm doi: 10.3917/reru.151.0027

Iannelli, F., Koher, A., Brockmann, D., Hövel, P., & Sokolov, I. M. (2017, jan). Effective distances for epidemics spreading on complex networks. *Physical review. E*, *95*(1-1), 12313. doi: 10.1103/PhysRevE.95.012313

Imai, M., Halfmann, P. J., Yamayoshi, S., Iwatsuki-Horimoto, K., Chiba, S., Watanabe, T., ... Kawaoka, Y. (2021, jul). Characterization of a new SARS-CoV-2 variant that emerged in Brazil. *Proceedings of the National Academy of Sciences of the United States of America*, *118*(27). doi: 10.1073/pnas.2106535118

Instituto Brasileiro de Geografia e Estatística. (2011). Demographic Census 2010. Retrieved from http://www.sidra.ibge.gov.br/cd/defaultcd2010.asp?o=4&i=P.

Jang, W., & Yao, X. (2011). Interpolating Spatial Interaction Data1: Interpolating Spatial Interaction Data. *Transactions in GIS*, *15*(4), 541–555. doi: 10.1111/j.1467-9671.2011.01273.x

Kraemer, M. U., Sadilek, A., Zhang, Q., Marchal, N. A., Tuli, G., Cohn, E. L., ... Brownstein, J. S. (2020). Mapping global variation in human mobility. *Nature Human Behaviour*, *4*(8), 800–810. Retrieved from http://dx.doi.org/10.1038/s41562-020-0875-0 doi: 10.1038/s41562-020-0875-0

Lai, S., Ruktanonchai, N. W., Zhou, L., Prosper, O., Luo, W., Floyd, J. R., ... Tatem, A. J. (2020). Effect of non-pharmaceutical interventions to contain COVID-19 in China. *Nature*, *585*(7825), 410–413. Retrieved from http://dx.doi.org/10.1038/s41586-020-2293-x doi: 10.1038/s41586-020-2293-x

Lam, N. S.-N. (1983, jan). Spatial Interpolation Methods: A Review. *The American Cartographer*, *10*(2), 129–150. Retrieved from https://doi.org/10.1559/152304083783914958 doi: 10.1559/152304083783914958

Landmann-Szwarcwald, C., & Macinko, J. (2016). A panorama of health inequalities in Brazil. *International Journal for Equity in Health*, 1–3. Retrieved from http://dx.doi.org/10.1186/s12939-016-0462-1 doi: 10.1186/s12939-016

-0462-1

Lee, W. D., Qian, M., & Schwanen, T. (2020). The association between socioeconomic status and mobility reductions in the early stage of England's COVID-19 pandemic. *Health Place*.

Levin, M. W., Shang, M., & Stern, R. (2021). Effects of short-term travel on COVID-19 spread: A novel SEIR model and case study in Minnesota. *PLOS ONE*, *16*(1), 1–16. Retrieved from https://doi.org/10.1371/journal.pone.0245919 doi: 10.1371/journal.pone.0245919

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Journal of Experimental Social Psychology Detecting outliers : Do not use standard deviation around the mean, use absolute deviation around the median. *Experimental Social Psychology*, 4–6.

Li, X., Tian, H., Lai, D., & Zhang, Z. (2011). Validation of the gravity model in predicting the global spread of influenza. *International Journal of Environmental Research and Public Health*, *8*(8), 3134–3143. doi: 10.3390/ijerph8083134

Li, Sabrina, Pereira, R., Prete, C., Zarebski, A., & Emanuel, L. (2020). Social and racial inequalities in COVID-19 risk of hospitalisation and death across São Paulo state, Brazil. , 1–23.

Lv, M., Luo, X., Estill, J., Liu, Y., Ren, M., Wang, J., ... Wang, X. (2020, apr). *Coronavirus disease (COVID-19): A scoping review* (Vol. 25) (No. 15). European Centre for Disease Prevention and Control (ECDC). Retrieved from /pmc/articles/PMC7175649/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC7175649/ doi: 10.2807/1560-7917.ES.2020.25.15.2000125

Malta, M., Murray, L., da Silva, C. M. F. P., & Strathdee, S. A. (2020). Coronavirus in Brazil: The heavy weight of inequality and unsound leadership. *EClinicalMedicine*, *25*, 100472. Retrieved from https://doi.org/10.1016/j.eclinm.2020.100472 doi: 10.1016/j.eclinm.2020.100472

Mascola, J. R., Graham, B. S., & Fauci, A. S. (2021, apr). SARS-CoV-2 Viral Variants—Tackling a Moving Target. *JAMA*, *325*(13), 1261–1262. Retrieved from https://doi.org/10.1001/jama.2021.2088 doi: 10.1001/jama.2021.2088

METRÔ-SP. (2017). Pesquisa Origem Destino 2017: A mobilidade urbana da Região Metopolitana de São Paulo em detalhes. , 136. Retrieved from http://www.metro.sp.gov.br/pesquisa-od/arquivos/Ebook%20Pesquisa%20OD%202017_final_240719_versao_4.pdf

Ministério da Economia/Fundação Instituto Brasileiro de Geografia e Estatística. (2019). RESOLUÇÃO No 3, DE 26 DE AGOSTO DE 2019. *DIÁRIO OFICIAL DA UNIÃO*. Retrieved from https://www.in.gov.br/web/dou/-/resolucao-n-3-de-26-de-agosto-de-2019-212912572

Moreno-Monroy, A. I., & Ramos, F. R. (2020). The impact of public transport expansions on informality: The case of the São Paulo Metropolitan Region. *Research in Transportation Economics*, 100928. Retrieved from https://www.sciencedirect.com/science/article/pii/S0739885920301268 doi: https://doi.org/10.1016/j.retrec.2020.100928

Neiva, M. B., Carvalho, I., Costa Filho, E. D. S., Barbosa-Junior, F.,

Bernardi, F. A., Sanches, T. L. M., ... Alves, D. (2020, oct). Brazil: the emerging epicenter of COVID-19 pandemic. *Revista da Sociedade Brasileira de Medicina Tropical*, *53*, e20200550–e20200550. Retrieved from https://pubmed.ncbi.nlm.nih.gov/33111917https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7580283/ doi: 10.1590/0037-8682-0550-2020

Oliver, N., Oliver, N., Lepri, B., Lepri, B., Sterly, H., Lambiotte, R., ... Vinck, P. (2020). Mobile phone data for informing public health actions across the COVID-19 pandemic life cycle. *Science Advances*, *6*(23), 1–7. doi: 10.1126/sciadv.abc0764

Peixoto, P. S., Marcondes, D., Peixoto, C., & 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 July), 1–23. doi: 10.1371/journal.pone.0235732

Pinto Neto, O., Kennedy, D. M., Reis, J. C., Wang, Y., Brizzi, A. C. B., Zambrano, G. J., ... Zângaro, R. A. (2021). Mathematical model of COVID-19 intervention scenarios for São Paulo—Brazil. *Nature Communications*, *12*(1), 418. Retrieved from https://doi.org/10.1038/s41467-020-20687-y doi: 10.1038/s41467-020-20687-y

Regmi, L. (2015, nov). An Overview of Population Growth Trends of Nepal. *Journal of Institute of Science and Technology*, *19*, 57. doi: 10.3126/jist.v19i1.13828

Ribeiro, H. V., Sunahara, A. S., Sutton, J., Perc, M., & Hanley, Q. S. (2020). City size and the spreading of COVID-19 in Brazil. *PLoS ONE*, *15*(9 September 2020), 1–12. doi: 10.1371/journal.pone.0239699

Rocha, R., Atun, R., Massuda, A., Rache, B., Spinola, P., Nunes, L., ... Castro, M. C. (2021). Effect of socioeconomic inequalities and vulnerabilities on health-system preparedness and response to COVID-19 in Brazil: a comprehensive analysis. *The Lancet Global Health*, *9*(6), e782–e792. doi: 10.1016/S2214-109X(21)00081-4

Savaris, R. F., Pumi, G., Dalzochio, J., & Kunst, R. (2021). Stay-at-home policy is a case of exception fallacy: an internet-based ecological study. *Scientific Reports*, *11*(1), 5313. Retrieved from https://doi.org/10.1038/s41598-021-84092-1 doi: 10.1038/s41598-021-84092-1

Schlosser, F., Maier, B. F., Jack, O., Hinrichs, D., Zachariae, A., & Brockmann, D. (2021). COVID-19 lockdown induces disease-mitigating structural changes in mobility networks. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(52), 32883–32890. doi: 10.1073/PNAS.2012326117

Schroeder, R. (2014). Big Data and the brave new world of social media research. *Big Data and Society*, *1*(2), 2053951714563194. Retrieved from https://doi.org/10.1177/2053951714563194 doi: 10.1177/2053951714563194

Secretaria da saúde do estado de São Paulo. (2020). *SIMI-SP: Pacientes internados por Síndrome Respiratória Aguda Grave (SRAG).* Retrieved from https://www.saopaulo.sp.gov.br/planosp/simi/

Singh, A., Yadav, A., & Rana, A. (2013). K-means with Three different Distance Metrics. *International Journal of Computer Applications*, *67*(10), 13–17. doi: 10.5120/11430-6785

The Lancet. (2020). COVID-19 in Brazil: "So what?". *The Lancet*, *395*(10235), 1461. Retrieved from https://www.sciencedirect.com/science/article/pii/S0140673620310953 doi: https://doi.org/10.1016/S0140-6736(20)31095-3

The Lancet Respiratory Medicine. (2020, dec). *COVID-19 transmission—up in the air* (Vol. 8) (No. 12). Lancet Publishing Group. Retrieved from https://www. doi: 10.1016/S2213-2600(20)30514-2

Tizzoni, M., Bajardi, P., Decuyper, A., Kon Kam King, G., Schneider, C. M., Blondel, V., ... Colizza, V. (2014). On the Use of Human Mobility Proxies for Modeling Epidemics. *PLoS Computational Biology*, *10*(7). doi: 10.1371/journal.pcbi.1003716

Vasconcellos, E. (2017, sep). Urban transport policies in Brazil: The creation of a discriminatory mobility system. *Journal of Transport Geography*, *67*. doi: 10.1016/j.jtrangeo.2017.08.014

Wen, T.-H., Lin, M.-H., & Fang, C.-T. (2012, sep). Population Movement and Vector-Borne Disease Transmission: Differentiating Spatial–Temporal Diffusion Patterns of Commuting and Noncommuting Dengue Cases. *Annals of the Association of American Geographers*, *102*(5), 1026–1037. Retrieved from https://doi.org/10.1080/00045608.2012.671130 doi: 10.1080/00045608.2012.671130

Wright, A. L., Sonin, K., Driscoll, J., & Wilson, J. (2020, dec). Poverty and economic dislocation reduce compliance with COVID-19 shelter-in-place protocols. *Journal of economic behavior organization*, *180*, 544–554. Retrieved from https://pubmed.ncbi.nlm.nih.gov/33100443https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7568053/ doi: 10.1016/j.jebo.2020.10.008

Yang, Y., Herrera, C., Eagle, N., & González, M. C. (2014). Limits of predictability in commuting flows in the absence of data for calibration. *Scientific Reports*, *4*, 1–9. doi: 10.1038/srep05662

Zarikas, V., Poulopoulos, S. G., Gareiou, Z., & Zervas, E. (2020). Clustering analysis of countries using the COVID-19 cases dataset. *Data in Brief*, *31*, 105787. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352340920306818 doi: https://doi.org/10.1016/j.dib.2020.105787