Analysis of the relationship between the diffusion of COVID-19 and the presence of metro, train and LRV in the neighborhoods of Rio de Janeiro

1 Introduction

1.1 Background

The COVID-19 has significantly impacted the whole world and Brazil is no exception. The country has already registered more than 14 million cases and 400 thousand deaths. Rio de Janeiro is one of the most impacted states of Brazil, especially the state's capital, the city of Rio de Janeiro. The city of 6.7 million habitants has already recorded 264 thousand cases of COVID-19 and 24 thousand deaths. Therefore, stopping the diffusion of COVID-19 in Rio is essential.

Many studies have showed a direct relationship between the use of public transportation, including metro, train, and light rail vehicles (LRV), and the diffusion of COVID-19 in urban areas. People are usually more exposed to the being infected during commutes in metro, train, and light rail vehicles (LRV), especially if the wagon is crowded. Although the use of these modes of transportation has significantly decreased in Rio since the beginning of the pandemic, they still carry more than 9 million passengers every month.

1.2 Problem

Metro, train and LRV might be vector for the diffusion of the COVID-19 through the city of Rio de Janeiro, as is the case in other cities, e.g., New York and Hong Kong. However, there are no studies on this topic in Rio. This project aims to test if the following hypothesis is true: metro, train, and LRV are contributing to the diffusion of COVID-19 in the City of Rio de Janeiro.

1.3 Interest

This information can inform public health officials in Rio about more effective ways of halting the diffusion of COVID-19 in the city. This is even more important now that new variants of the virus are starting to spread in Brazil. Besides, companies may use this result to decide which mode of transportation to recommend to their employees. Some companies in Brazil are already providing individual modes of transportation to their employees, for them to avoid public transportation. If the proposed hypothesis is confirmed, this will give this business more elements to justify this kind of policy. Besides, it may encourage other companies to adopt similar policies.

2 Data acquisition and cleaning

2.1 Data sources

The first source of data used is the register of COVID-19 cases made available by Rio de Janeiro government on the website https://coronavirus.rio. Data on the modes of transportation analyzed were based on the information available at the website's of the metro company (https://www.metrorio.com.br), the train company (https://www.supervia.com.br), and the LRV company (https://www.vltrio.com.br/#/). The lines and stations for each of these modes of transportation were identified in the websites and the geographic coordinates of each of the stations were collected using Google maps.

Besides, data about Rio de Janeiro 163 neighborhoods were retrieved from the Wikipedia. Data on the population of each of the neighborhoods was collected on the Rio de Janeiro government website: https://www.data.rio/search?collection=Dataset&tags=população. Finally, it was also necessary to access the Rio de Janeiro government website to retrieve a geojson file of Rio neighborhoods, which is used to create maps.

2.2 Data cleaning and use

First, I use the dataset of COVID-19 cases to compile a list of cases per month per neighborhood for the period that is analyzed: from January 2020 to April 2021. Then, I combine this data with the data on the population of each neighborhood to calculate the cases of COVID-19 per population for each of these neighborhoods for each month in the period analyzed. However, the population data is from the last Brazilian census, in 2010. The problem is that three neighborhood, Jabour, Lapa, and Vila Kennedy were created after this year. Therefore, I have to adjust the dataset of COVID-19 cases to combine these three neighborhoods with the ones they used to be part of: Senador Camará, Centro, and Bangu, respectively.

This same problem is detected in the dataset of Rio de Janeiro neighborhoods, retrieved from Wikipedia. Therefore, I do the same adjustment in this dataset. Combining these three datasets, I create a data frame of COVID-19 cases per month per neighborhood per population of the neighborhood. Besides, I also create a second data frame to store the data of cumulative COVID-19 cases per neighborhood per population of the neighborhood.

The next step is to prepare the geojson file to be used to create the maps. It is necessary to adjust the neighborhood names in the file and then attribute a number to each neighborhood. The necessary information to create the maps are retrieved from the geojson file into a data frame and combined with the numbers attributed to the neighborhoods. This data frame is combined with the data frame with the data of cumulative COVID-19 cases per neighborhood per population of the neighborhood to create a choropleth map to display the data with a time slider, as well as other analysis.

Then, I divide the dataset on the modes of transport in three sets, one for each mode. These sets are used to create a map to displaying these stations' locations. Finally, this map is combined with the choropleth map to create a third map. This final map displays the evolution of COVID-19 cases over time in each neighborhood and the modes of transportation present there.

Finally, the data frame of COVID-19 cases per population for each of these neighborhoods by month is combined with the information of the modes of transportation present in each neighborhood. This final data frame is used to perform a series of statical analysis to determine if the project hypothesis is valid or not.