Sidharth Mallela, Eric Ji
Civeng C88

Urban_Metro_Analysis (V1)

Introduction

Problem of Interest: Currently, São Paulo, Brazil, has an influx of bus routes but not enough metro routes. This is causing an extremely complex system where it probably takes much longer to commute to a location because there are not enough metro lines. Creating a map of more metro lines based on the best direction, etc… will create simplicity in Sao Paulo's transportation system. We are using two datasets containing the public transportation system of São Paulo, a major Brazilian city from Kaggle (https://www.kaggle.com/datasets/mateuscco/sao-paulo-transportation-service/data?select=routes.txt), (https://www.kaggle.com/datasets/mateuspicanco/sao-paulo-geospatial-features)

Pairing the routes.txt and frequencies.txt files, which "specifies the interval between departures in different periods of the day." We plan to understand which areas aren't optimal for transportation. We plan to use the methods of longest shortest path, clustering coefficient, density choropleth map, mean, standard deviation, quartiles, on bus routes, metro lines, and population. For Metro mapping, we used FOLIUM & BRANCA to create the maps, and simply mapped new metro lines to the most dense spots in the density choropleth map. We are planning to use the probability density function to see which areas require high demand. This will help us see where the problems and burdens lie for commuters. We will find the clustering coefficient and longest shortest path to see the areas with the least amount of connectivity. These methods can also help us examine the areas that are underdeveloped and will show us non-optimal routes that are inefficient and congested. This will help us see if there are any underlying issues beyond the scope of just transportation. After extracting route information from a transportation dataset, the nodes will represent stops that spread out from metro lines, and the edges will represent the routes. This network is fascinating because it depicts a city's transportation system, showing how various sites are connected by bus, rail, and metro lines, among other routes.

This Research Paper, similar to our network, is conducted by the Sao Paulo Metropolitan Society on the effectiveness of their metro lines. The main question to answer was whether the mapping of the current metro lines could be more optimized. This paper has some ideas we want to incorporate into the project in the future, including taking complete traffic data to account for heavy construction sites and data of the age groups in each prefecture.

For the presentation, we will display all of the methods for which we will have an analysis. Instead of presenting one singular holy grail fix for the transportation system in São Paulo, we will propose many different solutions considering the pros and cons of each. To support our solutions, we will use our visualization interactively so the audience can have a grasp on the transportation system of São Paulo.

Data and Methods

We used datasets pertaining to São Paulo's transportation services, such as those with regard to routes, stop times, calendar information, fare attributes, fare rules, frequencies, trips, stops, and shapes, to make up the input data. The'sao_paulo_data' directory contains CSV files that serve as the source of these datasets. We also used GEOJSON files to create the mapping of the prefecture boundaries. Route IDs, route kinds (e.g., Metro, Rail, Bus), route colors, service IDs, trip IDs, stop IDs, arrival/departure timings, fare IDs, and other relevant information are all included in each dataset. The datasets are organized tabularly, with rows denoting distinct records (such as routes, stops, and trips) and columns denoting these records' properties. The most important column would be POLYGON or Geometry values as that is what we use with Folium to create the mappings. Before beginning any analysis or visualization, it is imperative to do data preparation operations to guarantee the consistency and quality of the dataset, which is why we had an entire section of the code dedicated to setting up the Data Frames, and only then did we move forward with any visualizations. The main methods that we used for the metro mapping are FOLIUM and BRANCA, which are basically libraries that work in conjunction with GEOPANDAS to create these stunning maps. I had to merge the two datasets to include the POLYGON values from Sao_Paulo_Data and the Density Values parquet files. Grabbing the density values along with the prefecture names using pandas methods, we made connections from the most dense region (center of Sao Paulo) to other dense regions closer to the center. As you can see in the map, it avoided the majority of low density areas (below 350), which is our population density threshold.

For one of our methods, we found every stop location a metro line would encounter. Each stop location leads to a specific area, which almost all happen to have bus routes. According to our datasets (link above), there is an abundance of buses compared to metros, 1347 to 6, respectively. Because there are so many bus routes, it would be insightful to rank each metro line by the amount of connecting bus routes along its stops. Thus, we found the average clustering coefficient of bus routes with respect to the metro line, 6 in total. This can help us determine the efficiency of each metro line, helping us determine which line will need the most assistance. As an additional note, finding the average clustering coefficient for the metro lines themselves would not be beneficial because they are all separated from each other; if we had done so, we would have gotten a coefficient of 0. There are a total of 6 directed graphs, each with nodes representing the neighborhood name with that bus stop, and links representing a route to each node. Below is the number of nodes and links for each metro line

METRÔ_L5 Number of Nodes: 476
METRÔ_15 Number of Edges: 762
METRÔ_L1 Number of Nodes: 567
METRÔ_L1 Number of Edges: 936

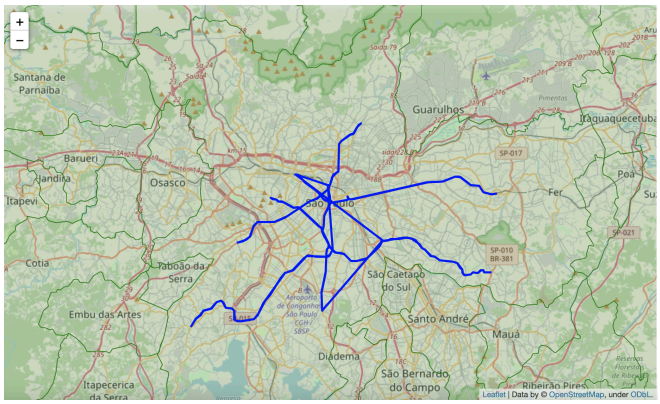
METRÔ_L2 Number of Nodes: 557
METRÔ_L2 Number of Edges: 915
METRÔ_L3 Number of Nodes: 515
METRÔ_L3 Number of Edges: 813
METRÔ_L4 Number of Nodes: 500
METRÔ_L4 Number of Edges: 817
METRÔ_L5 Number of Nodes: 376
METRÔ_L5 Number of Edges: 575

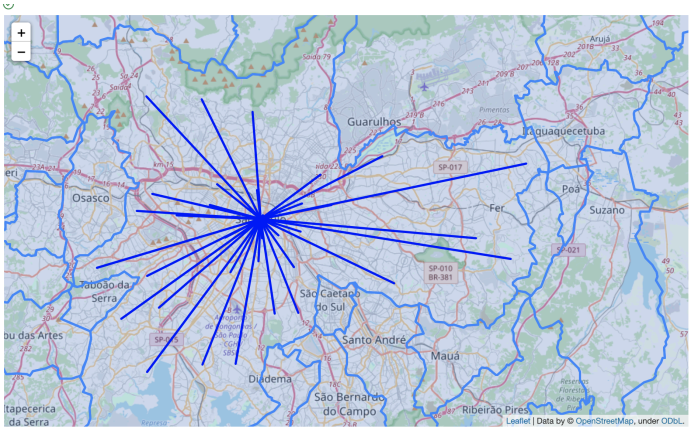
We will pair the method above with the method of finding the mean, standard deviation, quartiles, and max of bus routes, metro lines, and railways. This will help us quantify the importance of each transportation type which we will use in all the other methods.

For another method, we will be using the longest shortest path. We will find the shortest and longest path of the bus stops along each of the 6 metro lines. This method will help me determine which bus stop, which corresponds to its general area, has the least amount of connectivity. Because we are linking all of these six metro lines together, we will be able to start a scattering effect from each stop along the metro lines. By counting the number of appearances of each node in each of the longest shortest paths, we will be able to rank which neighborhood needs the most improvement. Adding metro lines in the direction of that area will most likely improve connectivity. Alongside this, we will pair another method of histogram visualization of indegree and outdegree to help us confirm our results and determine their accuracy.

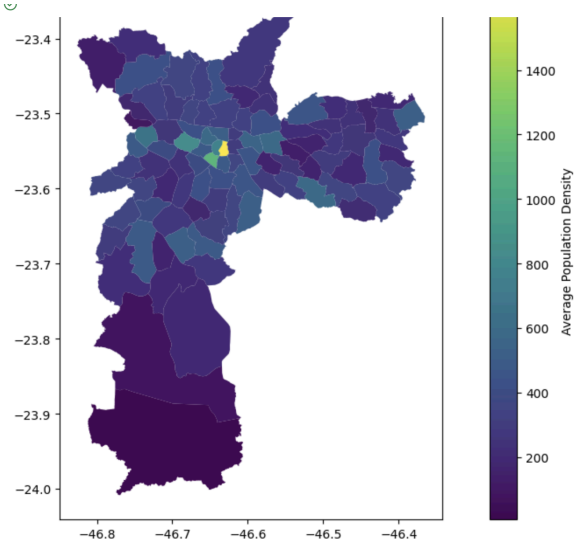
Analysis of the Results

Existing Folium Map of the Sao Paulo Metro Lines

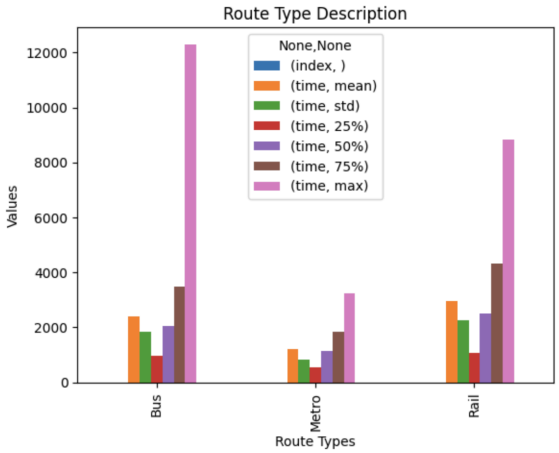




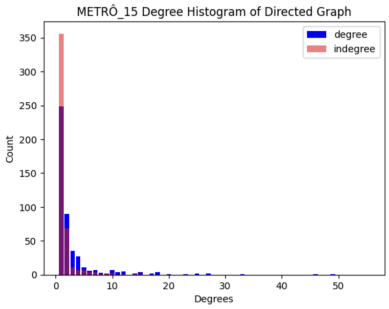
New Metro Line paths developed based off highly dense areas of Sao Paulo

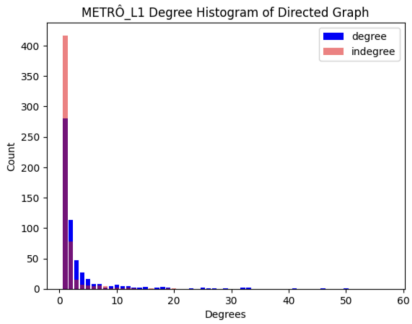


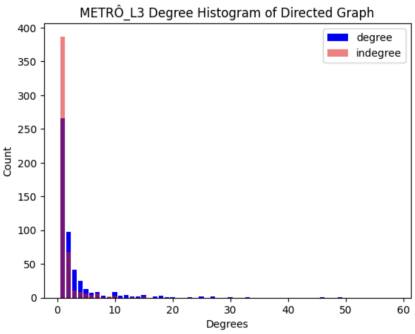
Density Choropleth map of Sao Paulo, most densely populated towards the center

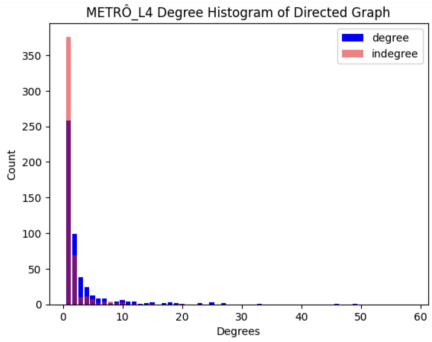


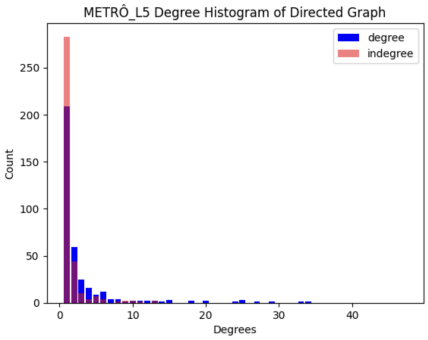
Above are the (mean, standard deviation, quartiles, and max) of bus routes, metro lines, and railways. Where values represent the time in seconds. Upon first glance, the most extreme observation is the max time for Bus Routes. The max route takes 3 hours and 20 minutes! Because the mean is less than 50% of the max time, this indicates that there are few areas that are very difficult to get to. Adding a metro line towards that direction could be useful as the metro's longest route time is only around 1 hour.











Out and in-degree distributions for all the bus routes with respect to each metro line. According to the histograms above, stops that connect along the metro line to other bus routes seem to usually have degrees less than 7. Most of the degrees, including indegree, are 1, signifying that there is only one connection between a lot of the nodes. Metro Line 5 (METRÔ_L5) looks the least connected because when the degree is 2, it has a count lower than all the other metros.

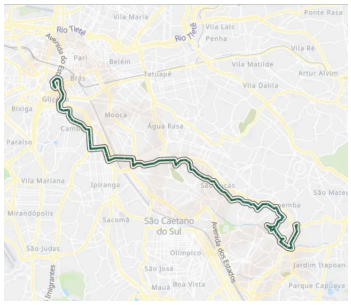
The picture above shows Metro Line 15 in gray and Metro Line 5 in purple
METRÔ_15 average_clustering: 0.04420570525777398
METRÔ_L1 average_clustering: 0.04037770803215402
METRÔ_L2 average_clustering: 0.040609384792402774
METRÔ_L3 average_clustering: 0.0396819505124095
METRÔ_L4 average_clustering: 0.04251234302641284
METRÔ_L5 average_clustering: 0.0334653675767835

According to the average clustering of each metro line, metro line 15 has the highest average clustering coefficient of 0.04420570525777398, while metro line 5 has the lowest average clustering coefficient of 0.0334653675767835. The average clustering coefficient tells us the degree to which nodes are clustered together. In this case, it shows the level of connectivity between bus routes along the metro line. Because metro line 5 has the lowest average coefficient, it suggests that it has less interconnectedness, meaning worse traveling accessibility for passengers. Without looking at any other data, this low coefficient indicates that the gray line in the Subway map above needs more improvement compared to the other lines, while the purple line needs the least improvement. However, when we look at the average population Density Choropleth map of Sao Paulo, the bottom left of the middle yellow block has the highest average population density compared to its neighboring districts. Metro 15, the purple lines also happens to spread towards the bottom left. This finding makes sense because, during the construction of the metro, engineers probably added more connectively towards the areas with higher average population density. Even though there is this tradeoff with every metro line, by observing the neighboring districts using the choropleth map, we see that the direction of metro line 15 has an average population density sub 200 while metro line 5 has an average population density of around 1200. Even though our findings suggest that Metro Line 5 has the most connectivity, it still needs the most improvement due to the sheer size of the population.

Longest Shortest Path:

Below is the longest and shortest path of bus stop areas following the path of one of six metro lines:
METRÔ_15_LSP = ['Term. Vl Prudente', 'Vl. Industrial', 'Metrô Ana Rosa', 'Jd. Selma', 'Term. áGua Espraiada', 'Term. Grajaú', 'Term. Pq. D. Pedro Ii', 'Jd. Danfer', 'Metrô Penha', 'Jd. São Francisco', 'Shop. Aricanduva', 'Metrô Tamanduateí', 'Pq. Sta. Madalena', 'Mooca']

Below is the count of bus stop area appearances in the longest and shortest path of all metro lines:
Bus Location Area With Occurences Of Longest Shortest Path between All Metros: {'Term. Vl Prudente': 4, 'Vl. Industrial': 5, 'Metrô Ana Rosa': 5, 'Jd. Selma': 5, 'Term. áGua Espraiada': 4, 'Term. Grajaú': 4, 'Term. Pq. D. Pedro Ii': 6, 'Jd. Danfer': 3, 'Metrô Penha': 3, 'Jd. São Francisco': 3, 'Shop. Aricanduva': 3, 'Metrô Tamanduateí': 1, 'Pq. Sta. Madalena': 1, 'Mooca': 1, 'Term. São Mateus': 1, 'Metrô Carrão': 1, 'Jd. Sta. Terezinha': 1, 'Term. Varginha': 1, 'Shop. Interlagos': 1, 'Jabaquara': 1, 'Vl. Guacuri': 1, 'Sto. Amaro': 2, 'Capão Redondo': 1, 'Term. João Dias': 1, 'Metrô Artur Alvim': 2, 'Pq. Savoy City': 2, 'Metrô Vl. Prudente': 2, 'Jd. São Paulo': 1, 'Metrô Vl. Matilde': 1, 'Itaim Paulista': 1, 'Metrô Bresser': 1, 'Jd. Itápolis': 1, 'Term. Sacomã': 1, 'Metrô Vergueiro': 1, 'Lgo. São Francisco': 1, 'Vl. São José': 1, 'Pinheiros': 1, 'Paraisópolis': 1, 'Vl. Gilda': 1, 'Term. Guarapiranga': 1, 'Term. Pinheiros': 1, 'Metrô Barra Funda': 1, 'Term. Pirituba': 1, 'Term. Casa Verde': 1, 'Pça. Do Correio': 1, 'Term. Sapopemba': 1, 'Pça. Almeida Jr.': 1, 'Conj. Teotônio Vilela': 1}



Above is a picture of Term. Pq. D. Pedro Ii
A bus stop that appears most frequently in the longest shortest path in all metro lines: Term. Pq. D. Pedro Ii (Occurring 6 times).

This means that D. Pedro Ii is the least accessible. No matter what metro line you take, it will be hard to get to, as it appears six times, and there are six metro lines. A metro line in this district would be optimal.

<u>Future Work and Conclusions</u>

In the future we want to incorporate OSM API, to create a mapping between actual metro stations. Right now, we mapped the metro stations from the middle of Sao Paulo (the most dense spot), so if I had the actual metro stations locations, the metro mappings would not be in a straight line. Our project went pretty well overall and we were able to develop many visualizations that helped showcase the density mapping as well. Sidharth M. was in charge of the metro mapping cells, and the density choropleth mapping + the data preparation, and the choosing of which libraries to use. Eric J. worked on all the histogram visualizations, clustering coefficients, route type time descriptions, longest shortest path, and the node routes of the metro lines. On a larger scale, we are also looking into creating metro line visualizations for the outer edges of sao paulo prefectures into more rural areas. This Research mentioned how metro lines in rural areas often succeed as much of the targeted population is older. So, mapping just 1 or 2 metro lines to the center station in Sao Paulo would create a more interconnected version of the city.

Show your work

Deepnote
Personal Github (Sid)