Finding relevant OD Data

Seeking OD data for this project first began with internet searches surrounding the following key words: “OD”, “Origin-Destination data”, “metro/subway/rail”, “ridership”. Other public transportation modes were initially considered to supplement the investigation, so “public transport” was also part of the search.

From limited results, papers that utilized origin-destination data were searched for, which often involved single-city case study analyses of public transit ridership. Data was requested from the authors of those papers to see if they could be used for this investigation. In addition, specific targeting of metropolitan cities with known smart-data tracking or metro systems that utilized card payment or equivalent was sought after. Publicly available data through government and public transit agencies for those targeted systems were also searched for, which yielded the greatest success in data collection for recent OD data.

Cleaning the data to create OD Flow Matrixes

After finding OD data and refining the pool to have systems with data at the timescale of at most annual OD trips, the data had to be cleaned and formatted into a matrix for our investigation.

The original source of data and the code to clean up the data for each system is as follows:

Bay Area Rapid Transit (BART)

Data was retrieved from <https://www.bart.gov/about/reports/ridership>. As the data collected is already in matrix form, no data cleaning was necessary before using it.

Washington DC

Data was retrieved from <https://planitmetro.com/2012/10/31/data-download-metrorail-ridership-by-origin-and-destination/>. The code to convert the data into matrix form is:



Queensland Rail

Data was retrieved from <https://www.data.qld.gov.au/dataset/go-card-transaction-data>. From here, the data was cleaned up using the following code:



London Underground

Data was retrieved from the Transport for London API database, which is an open data source: https://api-portal.tfl.gov.uk/. The code is as follows:



Singapore

Data was retrieved from the Singapore Land Transit Authority’s open source API platform, which utilizes Postman. The introductions are here: <https://datamall.lta.gov.sg/content/dam/datamall/datasets/LTA_DataMall_API_User_Guide.pdf>, and the dataset used for this investigation is titled “Passenger Volume by Origin Destination Train Stations”. From that data, the matrix was created by:



Creating Adjacency Matrixes for the Metro Systems

With the OD flow matrixes complete, adjacency matrixes had to be created for our investigation to reflect the physical structure of each system. With the station names/codes in the flow matrixes and based on available system map images of the five respective systems, the adjacency matrixes were created manually by inputting a “1” in a cell when two stations are shown to be next to each other in a given rail line, or a “0” when they are not. It’s important to note that the adjacency matrix reflects the system map during the year in which the flow OD data is for, and is not necessarily a representation of the current metro system.

With both the flow matrixes and adjacency matrixes, the data can be processed through the “main.py” file.