BUFN403 First Deliverable

Om Duggineni
University of Maryland
odo@terpmail.umd.edu

Benjamin Knight
University of Maryland
bknight5@umd.edu

Jack Doggett
University of Maryland
jdogget1@terpmail.umd.edu

March 7, 2025

1 Abstract

This project creates a novel visualization, providing analysis of transportation access - including public transportation, walkability, and driving - and the relationship it has on real estate and social impacts to specific neighborhoods in the Washington D.C. metro area. We will curate a diverse dataset including transportation access and economic indicators such as bank loans, income, and other economic factors. Public census and geographic datasets will be used to allow users to explore relationships between transportation access and key financial statistics with an emphasis on showing interactions between metro and bus connections within an area and indicators related to the economic activity in a neighborhood.

2 Data Gathering

Goal: Build a custom dataset consisting of the below data points, each collected at the census block level, for all census block groups in the region surrounding Philadelphia, PA. We picked this region in order to maximize the amount of data we would have available for the project.

Data points:

1. A measure of transit accessibility, calculated through determining how far a person can get, on average, from a randomly selected location within the region within a certain number of minutes. This will include how far a person can get by only and driving, and by only taking public transportation.

2. Income and housing data, from the US Census Bureau

3. Number and average quality of points of interest:

   (a) Schools and average school ratings

   (b) Restaurants and average food service ratings

   (c) POI data & associated scores

We will use this data to map how accessible a place is to public transportation in comparison to metrics describing regional economic activity, bank deposits/loans, economic data, and census/income data.

The specific datasets that we plan on using are described below:

2.1 Public Transit Accessibility Data

2.1.1 Background

Worldwide, public transport has been found to be a positive indicator for economic development, including prominant economic indicators like the housing market [1]. Metro systems, specifically, have a large body of research demonstrating positive economic impacts. Metro systems have been shown to increase mobility, boost industry and housing around stations, and increase property values within a large radius of each station [3]. However, metros are extremely capital intensive, limiting the scope of projects while having large costs on the communities and environments they are built in [4].

Areas better connected by metro generally have higher density, large transit oriented developments, and clusters of businesses leading to higher percentage of high income workers and larger and more connected markets. Bus lines have a wide range of ridership, with a mix of service workers and commuters. Bus lines have mixed economic benefits, often providing the only form of transportation for workers in marginalized communities, but also lacking the frequency, connectedness, and infrastructure to bring the same benefits as the metro.

Obtaining data on how well-connected a particular region is via public transportation networks is very important to our project, as understanding how well areas are connected through public transit in a standardized manner allows us to determine how this may affect the economic activity within the region. However, this is challenging, as any measurement of how accessible an area is using public transportation has to be reflective of how individual users perceive the convenience of the system. For example, if an area is well-connected through bus and train networks but navigating said network takes a long time due to slow, suboptimal routes or delays, then public transit may have a smaller economic impact on the area.

Additionally, since metro systems work in tandem with the other kinds of transportation infrastructure in an area, which means that calculating how navigable a region is with public transit also requires knowledge of the other types of transportation infrastructure in an area, including walking, bus, and biking infrastructure [7].

2.1.2 Approach

We plan on using a novel approach by using mapping data from routing applications in order to predict what the public transit accessibility of a particular area will look like. Two open-source projects will be of particular benefit in this case - OpenStreetMap and GTFS.

OpenStreetMap is a public dataset of mapping data built collaboratively as a community project. In addition to being quite accurate, especially for walking and biking paths, it is freely available online for anyone to download. Many projects exist that have the ability to work with this data, both for displaying maps and for GPS-based navigation apps. The dataset is also used by many big companies [8] for their own navigation apps. (As an example of a navigation app that uses exclusively OpenStreetMap (and GTFS, explained below) data, try Organic Maps)

OpenStreetMap data is available to download from a daily-updated mirror at https://download.geofabrik.de/, with subsections of the map focusing on different areas of the world, although we will likely be focusing on the area surrounding Maryland for our specific analysis.

GTFS, the General Transit Feed Specification, is an open standard that public transit providers use to provide real-time and historical data on their transit networks. Specifically, it contains information such as the location of transit stops and the timings of arrivals and current locations of buses and trains.

By using OpenStreetMap data for mapping/navigation and GTFS data for the transit system, we can, essentially, use a routing engine to determine the boundary of how far a particular user will be able to travel in a given period of time using the area's public transit system and road system, producing a plot called an isochrone plot. The area contained within an isochrone describes how many different locations a person can reach using the public transit or road systems, and is a good measure for how well-connected the surrounding area is through the transit system.

2.2 Census Data

The US Decennial Census provides demographic and socioeconomic statistics at the census block and block group level. We plan to get general income, demographic data, and housing data for each block group in the census dataset and use these as variables to compare the rest of our data against. This will be obtained from the website of the US Census Bureau or an equivalent organization.

2.3 Consumer Review data

We plan on using Yelp's Open Dataset in order to provide data on reviews for local businesses in the area. Yelp's dataset contains business reviews for

businesses in the area, as well as information about their category of business and a lot of metadata.

3 Potential Algorithmic/Data Challenges

The main challenges associated with this project will be (in order):

1. Finding and gaining access to the datasets which we want to integrate into our project.
2. Determining our locations of interest based on the data that we have access to, as well as what granularity we should use to measure them.
3. For each census block group, using a routing engine for transit accessibility using OpenStreetMap data to calculate 10, 30, and 60 minute isochrones for public transit and road travel separately.
4. Calculate the correlation between the six isochrone square mileages and various census block group statistics
5. Calculate the point of interests that fit in each of the six isochrones for each census block group.
6. Create a regression model to predict census block group statistics from the six isochrones, and the points of interests inside each isochrone (including type of point of interest, number of that type, and average ratings of that type).
7. Creating a visualization of the results and making it available online.

4 Data Analysis

After our group finishes collecting data, we plan on doing two types of analysis on the data. These are described below:

4.1 Correlation Analysis

dOur goal in the first stage of analysis will be to determine the association between public transit activity (calculated using isochrone size) and the economic factors associated with the census block group that we plan on studying.

Our input and output variables will be:

Input	Output
Size of Isochrone (driving)	Income
Size of Isochrone (public transit)	Housing Status
	Property values (if obtainable)

Generally, the goal of this step will be to see how public transit accessibility is correlated with lower or higher economic factors in an area. This would include,

for example, determining whether low-income, medium-income, or high-income neighborhoods are more or less likely to have good public transit systems within them.

4.2 Predictive Analysis

In the second stage of analysis, our goal will be to see if we can build a model to predict the average income and public transit accessibility of a region given what points of interest are accessible from the region using our isochrone-based analysis. In order to do this, we plan to:

1. Use Yelp's Open Dataset in order to locate a set of points of interest within range of each census block, and compute what categories of businesses are represented most within the region. Calculate the average ratings for each business.

2. Build a decision-tree-based model in order to predict the average income of the region based on the data obtained above. Use Shapely values in order to see whether the presence of different kinds of businesses within the reachable area is correlated to higher/lower incomes and public transit accessibility for people within the region.

The goal of this analysis is to further explore how people are affected by the businesses that they can reach from within the region in which they live. This includes exploring:

1. What businesses are more likely to be found in places with high/low public transit accessibility?

2. Which is more predictive of demographics - businesses reachable within an area by driving or businesses reachable within an area by public transit?

3. Within what radius / time distance do businesses tend to have the most impact on people's lives within a region? Do the impacts caused by businesses tend to be "local" (only reflected in small isochrones) or "widespread" (represented in larger isochrones as well)?

5 User Interface + Data Visualization

Our final product will be an interactive visualization that users can interact with to determine how various factors across these datasets contribute to the statistics of census block groups. Users will be able to view our data, as well as general trends and conclusions that we observed in the process of analyzing it. This user interface will be hosted online and will be publicly accessible.

We plan on using React (may change) to design the visualization and use mapping tools to show map data in a user-interactive way. We will also have static graphs and text describing the statistical analyses we have done. We aim

to host the final product online using GitHub Pages, where we will also include access to the datasets we used in our project.

6 Discussion

6.1 Intended Audience

Our intended audience are stakeholders in real estate and the communities in which we are modeling. Our dashboard will be public and designed to communicate our insights clearly and powerfully with all users. We are especially interested in needs within the real estate industry and welcome collaboration in our dashboard design process.

6.2 Potential Advisors

Our group is looking for advisors in real estate who can help us understand needs in the industry and guide our research. This may give us better insight on how to analyze the economic data we obtain.

Additionally, we may consider briefly meeting with someone involved in transportation research at UMD to understand how transit can affect economic development. Groups we are considering for this include the Maryland Transportation Institute and the CATT (Center for Advanced Transportation Technology) Lab.

7 Team Roles

It's possible that these roles may change in the future (this will be communicated clearly).

Member	Role
Om Duggineni	Data Cleaning - Obtaining Data, Initial Processing of data
Benjamin Knight	Combining Data Sources + Data Analysis
Jack Doggett	Distributed Compute + User Interface

References

[1] Dorantes, L. M., Paez, A., & Vassallo, J. M. (2011). Analysis of House Prices to Assess Economic Impacts of New Public Transport Infrastructure: Madrid Metro Line 12. Transportation Research Record, 2245(1), 962-978. https://doi.org/10.1177/13634607221107827

[2] General Transit Feed Specification. (n.d.). Retrieved March 7, 2025, from https://gtfs.org/

[3] Lin, D., Broere, W., & Cui, J. (2022). Metro systems and urban development: Impacts and implications. Tunnelling and underground space technology, 125, 104509.

[4] TIWARI, G. (2013). Metro Rail and the City: Derailing Public Transport. Economic and Political Weekly, 48(48), 65–76. http://www.jstor.org/stable/23528925

[5] OpenStreetMap. (n.d.). OpenStreetMap. Retrieved March 7, 2025, from https://www.openstreetmap.org/

[6] Summary of deposits. (2023, June 1). FDIC. https://www.fdic.gov/bank-financial-reports/summary-deposits

[7] Truong, R., Gkountouna, O., Pfoser, D., & Züfle, A. (2018). Towards a better understanding of public transportation traffic: A case study of the Washington, DC metro. Urban Science, 2(3), 65.

[8] Who uses OpenStreetMap? (n.d.). Retrieved March 7, 2025, from https://welcome.openstreetmap.org/about-osm-community/consumers/