The Effect of Transportation Infrastructure on Trip Patterns throughout Brooklyn, NY

Timothy Miller

Dept. of Geography and Environmental Science, Hunter College

GTECH 70200: Spatial Data Analysis

Professor Jochen Albrecht

December 17, 2022

Introduction

At an initial glance, the New York City subway system appears to provide exhaustive coverage to its citizens. Upon closer inspection, it becomes apparent that it is heavily biased towards serving the Central Business District (CBD) in Lower Manhattan. It is relatively easy to travel from the outer-boroughs to the CBD. However, it is much more difficult to travel between the other boroughs or even within the same borough.

To address this concern, the Regional Plan Association (RPA) has proposed the Triboro line- a passenger railroad that would operate along an existing freight route. The route would begin in the northeastern Bronx, wind through central Queens and Brooklyn, and terminate in southwestern Brooklyn. The full route is mapped in Figure 1.

The Triboro may provide benefits through two key pathways. First, for residents with access to cars, they will be given a compelling alternative to driving between their destinations. This may reduce air pollution and congestion throughout the region. Second, for residents without access to cars, they will be able to move more freely throughout the service area of the Triboro. This may spur economic activity as patrons and employees are better connected with services and jobs. However, before we can understand how the Triboro may interact with trip patterns, we need to understand how trip patterns and infrastructure interact generally.

Through three categories of statistical tests, we explore the relationships between trip counts and transportation infrastructure. The first test is spatial auto-correlation. It examines the propensity for destinations to cluster in the same area. The second test is Ordinary Least Squares (OLS) regression. It examines how the variation in number of commutes between two areas can be explained by the amount of time it would take to travel between these two areas. The spatial auto-correlation and OLS tests both examine travel time determined by walking, driving, and taking the subway. The final test is network auto-correlation. It examines propensity for commutes to cluster around commutes with the same value.

Each of these tests explores a different dimension of commute data. Together, they provide interesting insights into general patterns for transportation infrastructure and trip patterns. With these insights, we can glean the potential role of the Triboro.

Conceptual Model

The total number of trips between two locations is driven by the desire to move from the origin to the destination and the ability to complete the travel. This travel desire may spring from several sources- from visiting friends to shopping for groceries. There are several data sources that attempt to capture this data, including the NYC mobility survey. Unfortunately, these data are not necessarily voluminous or veracious enough to generate reliable results. For this reason, we focus on The Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES). It exclusively describes trips between home and work. However, it is based

on high fidelity employment data. With this, we create a simplified conceptual model around work commutes.

For commutes, a propensity for a location to serve as an origin is related to the abundance of housing at that location. The propensity for a location to serve as a destination is related primarily to the amount of employment at that location. The number of desired trips that are actually completed is related to the amount of time that it would take to complete the trip. Trips that are shorter than fifteen minutes are not of interest because they can be completed by walking or biking. Trips that would take longer than sixty minutes are generally not taken regularly. The majority of trips take thirty minutes to complete. The number of trips will decrease as the travel time decreases from thirty minutes to fifteen minutes. They will also decrease as the travel time increases from thirty minutes to sixty minutes.

The two most expeditious intra-city transportation methods are highways and subways. Commuters must generally choose between these two methods of travel. Within US culture, commuters generally prefer cars as their first mode of travel. They will switch to the subway when it becomes more practical than car travel. If neither highways nor subways are practical, commuters will generally not make a trip regularly. The number of trips taken by highway are influenced by the routes and congestion levels of highways, parking availability at the origin, and parking availability at the destination. The number of trips taken by subway are influenced by the routes of subways, the proximity of a subway station to the origin, and the proximity of a subway station to the destination.

Highways and parking lots can accommodate significantly fewer trips than subways and subway stations. Several desired trips are not being made because the car infrastructure has met its capacity. The Triboro can be modeled as a subway line. It may increase the number of trips made between several origins and destinations by reducing the amount of time the trip would take by subway. It may also decrease the number of car trips as drivers move away from congested highways. This conceptual model is visualized in Figure 2.

Study Data

The conceptual model provides a guideline for the data required for our study. The LODES data is already established as a meaningful and practical data source. We use the 2019 NYS All Jobs data (download ny_od_main_JT00_2019.csv.gz). The conceptual model also calls for parking lot data. However, parking lots are often privately controlled and utilization data are impractical to obtain.

The conceptual model also guided the resolution used for commutes. LODES data are provided at the block level. However, the research question relates to trips that require the highway or subway system. Census blocks for NYC are generally small enough that people can walk between them. Because the research question is not interested in these trips, we can consolidate

these census blocks into larger areas. NYC collects census blocks and tracts into Neighborhood Tabulation Areas (NTAs). NTAs generally align with neighborhood boundaries within the city. They are mapped in Figure 3. They have an estimated average radius of 3500 meters. NTAs generally represent the minimum distance for trips that would benefit from the subway or highway system. They are used for the resolution of the analysis.

Data on congestion levels for highways are available through the Google Directions API. It generates a traffic model based on previous observations. It is also provides transit directions that account for proximity to subway stations. Three classes of requests were made to the Directions API. For each class of requests, a different mode of transportation is requested to make the trip. In addition to driving and subway directions, walking directions serve as a comparison baseline for driving and subway directions. Directions were found for every single pair of NTAs. The directions were considered undirected- the significant characteristics of the trips are the same in each direction. Each trip was given a departure time of Nov 21, 9:00am EST. The driving time used the "best guess" Google traffic model. The subway directions were constructed to request a transit route with subway as the preferred route. The travel time for each transportation mode was saved to a csv file. For the subway trips, the number of subway lines required to make the trip is also recorded. Each route starts and ends in a point in the NTA that is found using the R package “sf” 'point on surface' algorithm. This results in points that are roughly centered and guaranteed to be in the NTA. The requests were made with a Node.js script.

Maps throughout the report are generated from geographic data available from the NYC Open Data portal. This includes Census Tract and NTA equivalents, NTA boundaries, Borough Boundaries, Subway Lines, and Arterials & Major Streets.

Selection of Study Area

As part of the exploratory data analysis process, we mapped the number of jobs in each NTA throughout the four boroughs with access to the subway system. An NTA’s job count is calculated from the sum of all qualifying trips that list the NTA as its destination. In order for the trip to qualify, it must begin and end in the study area.

Midtown Manhattan was expected to have the greatest number of trips. However, we did not anticipate the extent of the disparity between Midtown Manhattan and the rest of the boroughs. Figure 4 shows that Midtown Manhattan has tens and even hundreds of thousands more jobs than other NTAs. Figure 5 shows an overwhelming proportion of commutes end in Midtown Manhattan. This disparity would likely drown out the transportation effects for other neighborhoods. We attempt to detect more subtle patterns by narrowing the study area to a single borough. As part of this process, we explore the intra-borough travel patterns.

To help select the borough to focus on, we examined trips that occur within each borough. From Figure 6 and 7, Manhattan is still dominated by commutes to Midtown. However, the Bronx, Brooklyn, and Queens trips are more evenly distributed. Looking at the total number of trips, Brooklyn has the most outside of Manhattan. Manhattan has 508,873, the Bronx has 114,719, Brooklyn has 370,256, and Queens has 265,386. As Brooklyn has the second most trips and these trips are distributed throughout the borough, it will be used as the sole borough of interest for the rest of the project.

Brooklyn has a number of parks. NYC lists these parks as a single NTA ("BK99") which is dispersed geographically across Brooklyn. Because these parks are neither a single geographic entity nor a major driver of trips, they are removed from analysis. This leaves 50 NTAs of interest.

Spatial Auto-correlation

The first statistical model is spatial auto-correlation. Spatial auto-correlation measures the propensity for areas with similar values to cluster together. In this case, the number of jobs in an NTA is the value of interest. From Figure 10, we can see that each NTA is a different area. To account for this variation, we standardized the job counts by dividing them by the area of their NTA. The resulting measurement is Jobs per km^2. We plot the frequency of occurrence for job counts in Figure 8. Examining this histogram, we see that Job count is skewed to the right. Moran's I, the test of spatial auto-correlation, assumes a normal distribution for the value of interest. To achieve this distribution, job count is transformed with the natural log. Figure 9 shows this transformation resulted in a more normal distribution. The spatial distribution of transformed job counts is visible in Figure 10.

Spatial auto-correlation also requires a definition for the weights of the neighbor relationships. We use four distinct neighbor weight definition. The first definition is the queen contiguity, in which two NTAs are neighbors if their borders physically touch. The final three definitions are related to the amount of time it would take to travel between the NTAs. Each definition relates to the travel method- subway, walking, or driving. We also transform the time values. First, increasing time generally decreases proximity. However, we want low times to correspond to high neighbor weights. Consequently, we take the inverse of time. Second, there is a non-linear relationship between time and job count. As part of the OLS analysis shown later in the report (e.g. Figure 20), we found that raising time to the power of one-eighth created a reasonably linear relationship between time and job count. This applied across walking, driving, and subway times. Consequently, $1 / seconds^{1/8}$ is used as the transformation for times entered into the neighbor weight matrix.

When finding directions for subway based trips, the Google Directions API returned 77 trips that failed to utilize the subway system; it was more practical to walk directly between the NTAs. This is not merely because the NTAs are physically close together. Some of these walking trips

are 40 minutes long- long enough that the subway would be preferable, if available. In the context of spatial auto-correlation, these NTAs are not neighbors. In the neighborhood matrix, they are given a weight of zero.

The Global Moran's I for each neighbor definition is available in Table 1. The Global Moran plots for each neighbor definition are available in Figures 11, 12, 13, and 14. We can see the Queen's contiguity has the greatest Moran's I at 0.39. This suggests a notable portion of jobs are located in close proximity to other jobs. Driving and Walking also have statistically significant Moran's I. However, they are not much different from zero. The Moran's I for Subway is not statistically significant.

We also performed a Local Indicator of Spatial Auto-correlation (LISA) test. The maps of the clusters and outliers are available in Figure 15. The first letter of the significance code refers to the value of the NTA. An H (High) indicates the job count of the NTA is above the mean. An L (Low) indicates it is below the mean. The second letter refers to the values of the NTA's neighbors. An NTA labeled HH has a value above the mean and its neighbor's values are generally higher than the mean.

The top right map of Figure 15 shows the LISA map for driving. We can see the eastern section of Brooklyn has a cluster of LL NTAs. This indicates there are few jobs within these tracts. Additionally, these tracts are isolated from NTAs with high levels of jobs. If an NTA is disconnected from High job count NTAs, then it may suffer economically. The bottom right map shows the LISA for subway. The same NTAs that are LL in the driving map are either Insignificant or LH in the subway map. This indicates that the subway system is able to connect each of the Low job count NTAs with some High job count NTA. It is able to bring residents of the Low job count NTAs to High job count NTAs. Additionally, these two maps are overlayed with their respective infrastructure. From a quick visual inspection, we can see the driving map has more miles of infrastructure. This indicates the subway system better serves Brooklyn residents while using fewer miles of infrastructure.

Table 1. Global Moran's I for neighborhood weights based on Queens Contiguity, Subway Times, Driving Times, and Walking Times

Parameter	Queens	Subway	Driving	Walking
Moran I	0.39	-0.045	-0.010	-0.009
Standard deviate	4.47	-3.53	6.62	6.21
P-value	3.9e-6	0.9998	1.8e-11	2.7e-10

Ordinary Least Squares

The ordinary least squares determines how much of the variation in the dependent variable can be explained by the independent variable and how much the dependent variable changes with the independent variable. In this instance, the independent variable is the amount of time it takes to travel between two NTAs using the infrastructure of interest. The dependent variable is the number of commutes that are accomplished between two NTAs.

For simplicity, a commute is considered undirected- the commute from NTA A to NTA B is equivalent to the commute from NTA B to NTA A. Similar to job count, we standardize commutes by the area. Because there are two NTAs involved in a commute, we take the sum of the origin and destination NTAs areas as the total area. The frequency of commute counts is shown in Figure 16. Similar to job count, it is skewed to the right. We apply a natural log transformation. This normalizes the data (Figure 17).

We also need to normalize the independent variable. First, we need to account for the 77 trips that are could not be accomplished via the subway. As this behavior is undefined, we remove the 77 routes from all three models. This leaves 1148 routes. Next, we apply the same transformations to time for the OLS that we applied for the spatial auto-correlation. Figures 19, 22, and 25 show the shape of the data before these transformation. Figures 20, 23, and 26 show the result of the transformations. Figures 21, 24, and 27 show the diagnostic plots for each model. Some non-linearity remains in the scatter plots. Additionally, some heteroscedasticity remains in the residuals of each linear model. However, the trends are not strong enough to undermine the results.

Each transportation method uses the same form of the model. The model equation is shown below. Every model is statistically significant and resulted in unique parameters. From Table 2, Walking has the highest R^2 of 0.370. Subway is close behind with a value of 0.350. Driving has the lowest R^2 with a value of 0.232. Subway also has the greatest coefficient for the time variable, with a value of 26.97. In contrast, driving time has a coefficient of 19.03. This means that decreasing the subway travel time between two points would lead to a greater increase in commutes when compared to decreasing the roadwork travel time. We can see this in Figures 28 and 29.

Table 3 helps illustrate the relationship between travel times and predicted commute counts. We can see that Subway trips of 50 minutes are generally associated with 2.7 commutes per km^2. This is statistically equivalent to the 2.9 commutes per km^2 that driving commutes of 25 minutes achieve. Commuters tend to tolerate higher subway travel times. Interestingly, the conceptual model predicted a non-monotonic relationship- commute count first increasing until reaching 30 minutes and then decreasing. However, the data reflects a monotonic relationship- commute counts consistently decrease as commute time increases.

Model equation: $Commutes\ per\ km^2 = exp(b_0 + b_1 / seconds^{1/8} + Error)$

Table 2. Values for parameters of Subway, Driving, and Walking models

Parameter	Subway	Driving	Walking
b_0 (95% CI)	-8.93 +/- 0.80	-6.57 +/- 0.82	7.01 +/- 0.62
b_1 (95% CI)	26.97 +/- 2.17	19.03 +/- 2.04	23.69 +/- 1.83
R^2	0.350, p = 2.2e-16	0.232, p = 2.2e-16	0.370, p = 2.2e-16
Error of the estimate	N(0, 0.67)	N(0, 0.73)	N(0, 0.66)

Table 3. Predicted commutes per km^2 for selected times in Subway, Driving, and Walking models

Minutes	Subway	Driving	Walking
12	18.5 +/- 1.3	6.0 +/- 1.5	29.8 +/- 1.3
25	6.6 +/- 1.3	2.9 +/- 1.5	12.0 +/- 1.3
50	2.7 +/- 1.3	1.5 +/- 1.5	5.4 +/- 1.3

Network Auto-correlation

In this instance, network auto-correlation examines the propensity for commutes with similar values to cluster together. For our model, each commute is a node. Two commute nodes are connected if they share an NTA in common. Figures 30 through 33 demonstrate the creation of a network auto-correlation neighborhood matrix. First, start with the NTAs as they exist in geographical space (Figure 30). Then, graph the commutes that connect them (Figure 31). From there, create an adjacency matrix to show the connections between NTAs (Figure 32). Finally, the neighborhood matrix is created by defining the commutes as the nodes (Figure 33). The commute neighbors are defined as those commutes which share an NTA. For the Moran's I analysis, we will use the weights from the neighborhood matrix that are conceptually represented by Figure 33. We also use the number of times a commute is made as the value of interest.

The Global Moran's I for the network auto-correlation of commutes is 0.25, $p < 2.2e\text{-}16$. The Global Moran Plot is shown in Figure 34. This indicates that commutes tend to cluster around commutes with the same count. However, the effect is not as strong as the queen contiguity for job counts. This makes intuitive sense. For the queen contiguity, high job value NTAs can only

be neighbored by the NTAs that physically border them. For the commute auto-correlation, commutes can be neighbored by any commute that shares an NTA with it. Also, high job count NTAs will attract both high and low count commutes. This may dilute the effect of high count commutes attracting each other. Despite these diluting effects, we still observe significant clustering.

Figure 35 displays the LISA of the commutes in space. For most of the maps, it is difficult to discern a pattern. It appears that High Low commutes (middle right map) forms clusters along the north and south of the map. Though, it is uncertain what this indicates.

Remembering the distribution of commutes from Figure 16, we know there are a few commutes of high value. Focusing on these top commutes may isolate interesting patterns. Figure 16 shows the network of commutes that are in the top 5% of counts (62 commutes total). The size of each node indicates its count of commutes. Applying a fast and greedy clustering algorithm, we can see the network can be separated into three clusters. Each node is colored based on the cluster it falls into. The map in the left of Figure 37 shows the cluster type for these top commutes. They mostly neighbor other high count commutes. The map in the right of Figure 37 shows which cluster each commute falls into. It appears the clusters are separated into the northern, central, and southern areas of Brooklyn. Additionally, cluster 1 mostly radiates out of a central NTA. This largely follows the pattern of the current Subway network. Returning to Figure 1, The Triboro could help provide a missing connection. Though it may miss cluster 2, it could connect the radials of cluster 1. This connection may also carry through to cluster 3.

Conclusion

The three statistical tests each revealed a different dimension of the transportation dynamic. First, the spatial auto-correlation revealed that the subway is a primary factor in connecting low job NTAs to high job NTAs. It also showed there is a significant propensity for jobs to cluster into contiguous neighborhoods. These factors should not be separated. Any effort to increase commutes between two NTAs should account for both subway infrastructure and any additional factors that makes clusters of neighborhoods attractive to jobs. These additional attractive factors are outside the scope of this study.

Next, the OLS demonstrated that subway infrastructure explains much of the variation in the number of commutes between two NTAs. If looking to increase the amount of commutes between two NTAs, a decrease in subway commute time will have a greater effect than a decrease in driving time.

Finally, the network auto-correlation revealed clusters of commutes that are isolated from each other. The Triboro's route would serve to connect much of these clusters. Paired with the evidence that subways are generally strong generators of trips, The Triboro should meaningfully change Brooklyn. This change would likely extend to Queens and the Bronx.