Citibike Station Use

STAT 787

Vitaly Druker

December 20th, 2017

Contents

1 Introduction

The CitiBike project (Citibike, n.d.) is a bike sharing program started in NYC in 2014. It has been considered successful and continues to grow every few years by adding new stations to existing neighborhoods and expanding to new ones.

There are many different factors that account for where Citibike stations should be placed. One prevailing theory is that they should be placed near subways so that commuters are able to piggyback their bike trip off of other forms of transit. (NACTO, n.d.)

Another theory brought forward in this study is that it's better to place CitiBike stations in a high concentration. The theory behind this is that more stations allow riders to have more choices in stations in case a particular station is out of bikes.

The report below will use spatial analysis and regression techniques to test both of these statements.

2 Data Sources

2.1 Citibike Data

CitiBike data is available online for free download (Citibike, n.d.). Data is available for each ride taken and includes start/stop times, start/stop stations along with some basic ridership information. A small sample of the data is shown below:

```
Observations: 9,884,307
Variables: 11
$ tripduration     <dbl> 997, 1904, 305, 250, 464, 1118, 394, 1449, 42...
$ starttime        <dttm> 2013-07-01 06:00:16, 2013-07-01 06:00:30, 20...
```

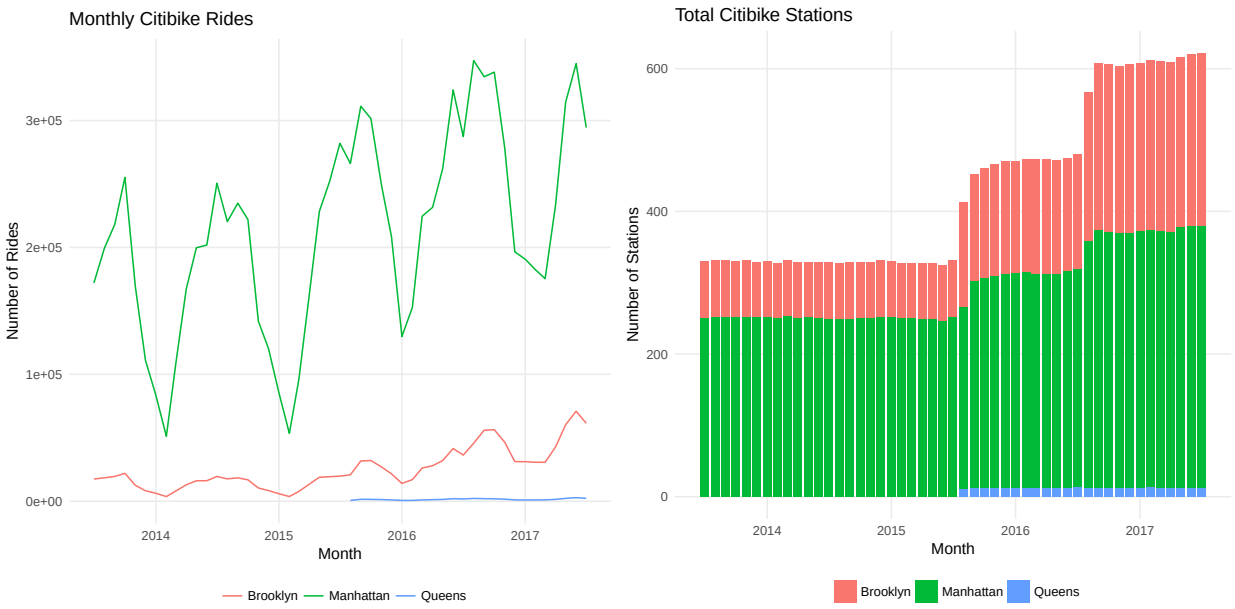

Figure 1: The two graphs above show general information about CitiBike ridership. *Left*: The graph shows monthly rides by NYC borough. *Right*: The number of unique stations in a given month broken out by borough.

```
$ stoptime         <dttm> 2013-07-01 06:16:53, 2013-07-01 06:32:14, 20...
$ startstationid   <chr> "436", "294", "385", "271", "477", "488", "30...
$ startstationname <chr> "Hancock St & Bedford Ave", "Washington Squar...
$ endstationid     <chr> "467", "375", "440", "390", "522", "497", "32...
$ endstationname   <chr> "Dean St & 4 Ave", "Mercer St & Bleecker St",...
$ bikeid           <dbl> 16199, 20281, 18143, 16370, 15497, 15502, 161...
$ usertype         <chr> "Subscriber", "Subscriber", "Subscriber", "Su...
$ birthyear        <dbl> 1979, 1949, 1988, 1962, 1975, 1957, 1963, 195...
$ gender           <chr> "2", "1", "1", "1", "1", "1", "2", "1", "1", ...
```

While the data is extensive, this report will only look at a single day worth of data. Nevertheless some graphs described below show some interesting views of the data to help the reader visualize the general feel of the data.

Figure 1 shows some general ridership statistics for CitiBike usage over time. The figure on the left demonstrates both seasonal variability in ridership along with a steady increase in general ridership. The figure on the right shows that the number of stations per station has not increased linearly, but happens in spurts and generally coincides with breaking out into a new neighborhood.

Figure 2 echoes much of what was shown in Figure 1, but plotted on a map for a few select days. Ridership is lower in the winter, and there was a clear expansion the 1st and 2nd panel, along with between the 3rd and 4th panel.

Station density data will be estimated by looking at the number of stations within 0.5 kms, which can be used as a good approximation of how far people are willing to walk. Figure 3 shows the distribution of the number of nearby stations.

2.2 Subway Entrance Data

Subway Entrance Data was used to estimate the number of close subway stations along with the distance to the nearest station. The analysis is the same as what was done for identifying the closest CitiBike stations.

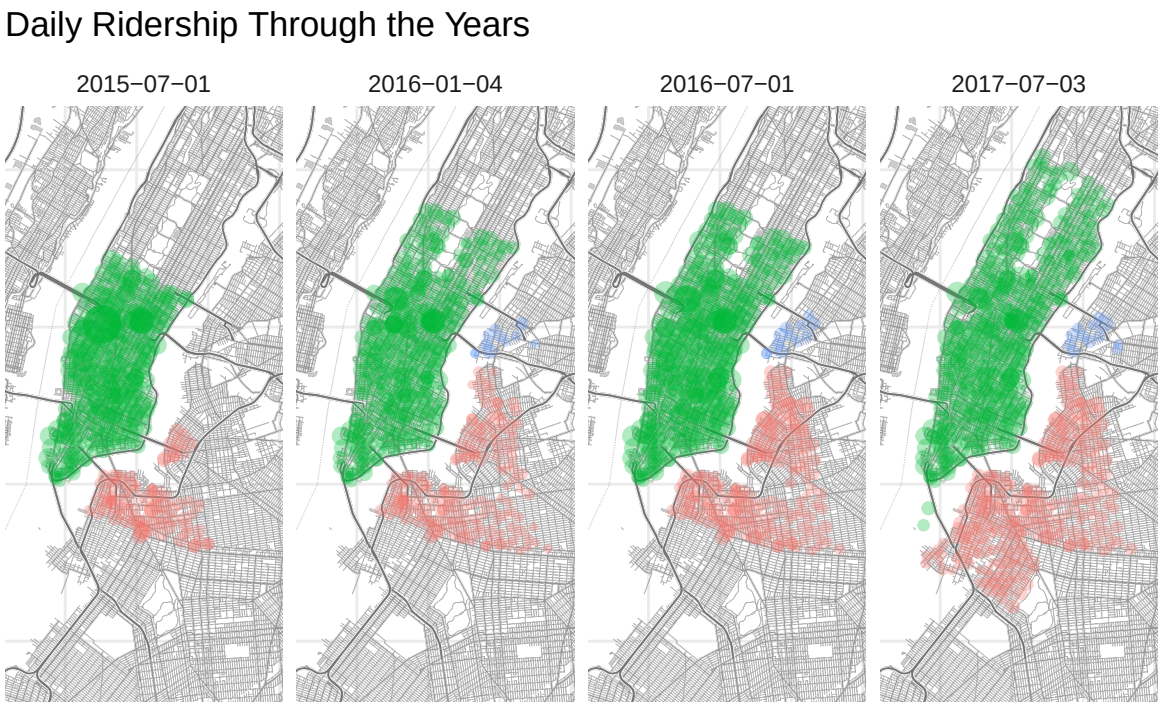

Figure 2: The figure above shows 4 days of CitiBike usage. Each dot represents a station and the size of each dot is proportional to the number of rides taken from that station.

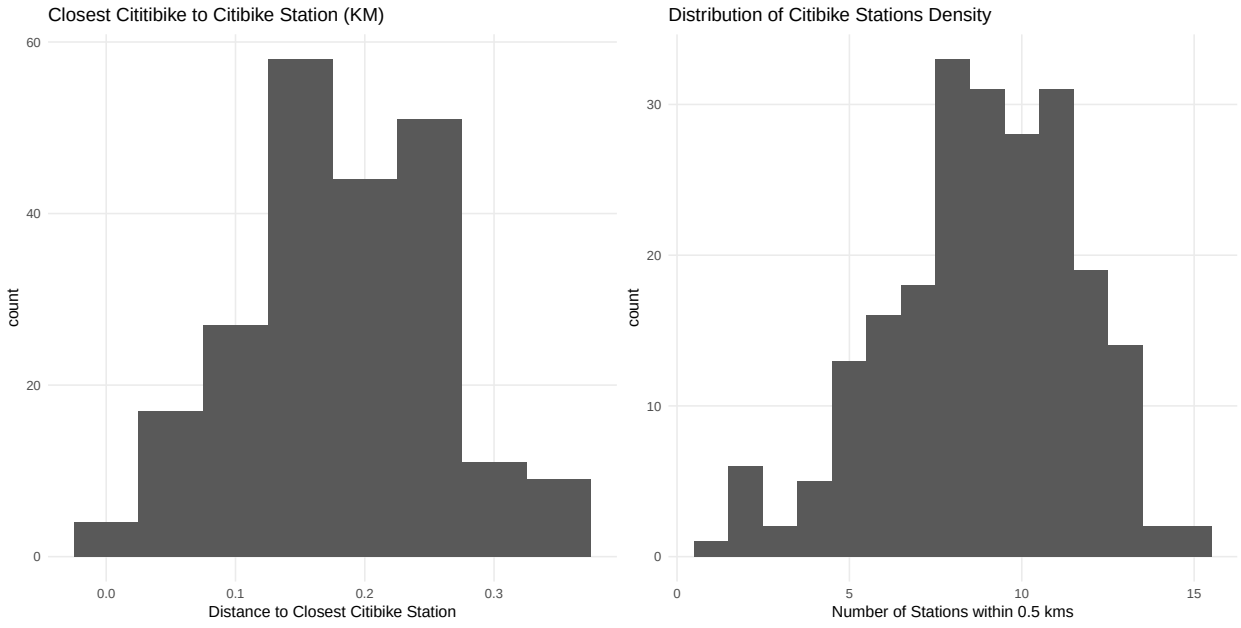

Figure 3: Histogram of number of close stations.

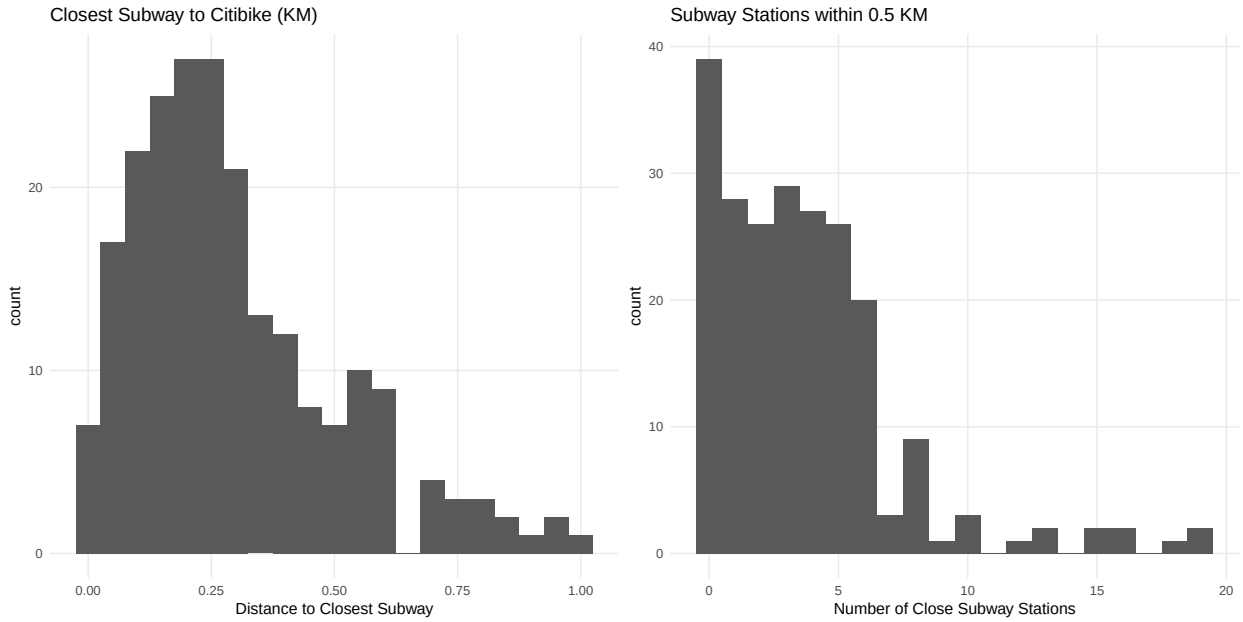

Figure 4: A histogram of the distance to the nearest subway, along with the number of stations within 0.5 kms.

Data was pulled from the MTA website (MTA, n.d.). Figure 4 shows similar data to that above.

3 Main Analysis

The main analysis will attempt to fit variograms to a single day of data. The outcome variable will be the number of rides taken between the hours of 6 am and 10 am during weekdays. The analysis will only include stations in lower Manhattan as it's a more homogeneous population.

3.1 Regression Analysis

Figure 5 shows a view of the data with subway stations and CitiBike stations. It's clear that there are some areas with high density of stations of either kind.

The model was fit using a linear model. Step wise model selection was used. The initial model was defined as:

$$log(n+1) = (Intercept) + \beta_{ClosestSubway} + \beta_{CloseSubways} + \beta_{CloesetCiti} + \beta_{CloseCitis}$$

where n is the number of rides taken at the station. 1 was added to deal with stations that had 0 rides.

The final model (after step wise selection) is shown below:

term	estimate	std.error	statistic	p.value
(Intercept)	3.64	0.05782	62.96	7.787e-128
close_subways	-0.04174	0.01075	-3.884	0.0001424

Residuals were checked and found to approximately normal and well dispersed. Only the close subways variables was left, and it had a negative effect on ridership. This analysis was repeated with variogram modeling.

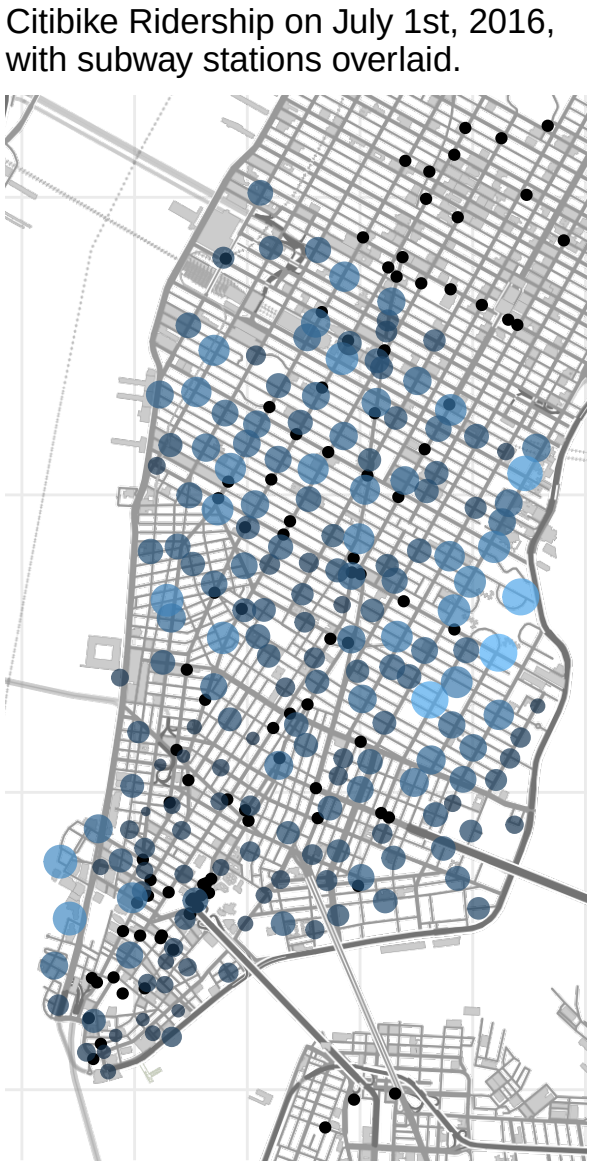

Figure 5: Rider ship is denoted at CitiBike stations with blue circles that vary in size based on the number of rides. Dark circles note subway entrances.

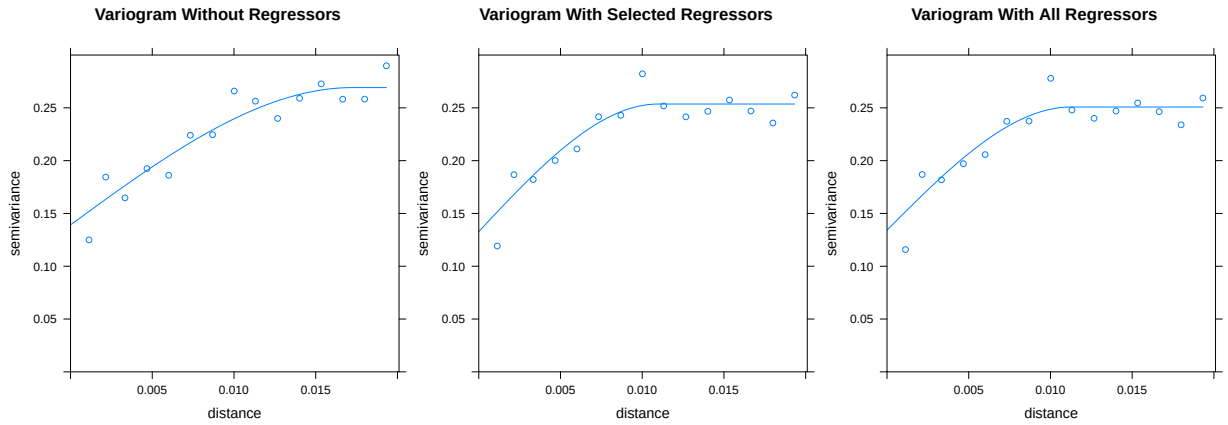

Figure 6: The two variorums are shown above. The scale of the distance can be multiplied by 100 to approximate kilometers. e.g. .015 on the scale is approximately 1.5 km. *Left*: The variogram fit without regressors. *Middle*: The variogram fit with the number of close_subways regressors. *Right*: All regressors.

3.2 Fitting the Variogram

A spherical variogram with a 0.01 nugget was fit to the data without regressors and then to same data but with the close_subways variable.

4 Discussion

Using regular regression, it appears that the number of close subways is the only variable that is significantly associated with outcome. The simple linear regression suggests that each subway drops rider usage by about 4% ($e^{-0.04} \approx .96$). This disagrees with the original citation that suggest that higher subway density would result in more ridership. Likewise, measures of CitiBike station density were not found to be significant in predicting ridership turnout.

Fitting the variogram on the entire data set produced a satisfactory fit. This suggests that there is a spatial relationship to the number of rides. Adding the regressors change the range property of the final variogram. A summary table of the two variograms (left and middle from figure 6).

Regression	model	psill	range
None	Nug	0.1392	0
None	Sph	0.13	0.01725
Selected	Nug	0.1329	0
Selected	Sph	0.1207	0.01097

There appears to be a small difference in the partial sill of the Spherical component, but the range change much more dramatically. What this suggests is that adding the closest subway regressor localized the cluster differences more than they were previously. Instead of the clusters reaching to about 1.7 km in distance, now they only go out to about 1 km. (Range should be multiplied by 100 to get to km scale.)

Further analysis can be made by expanding the number of regressors used to include other features mentioned in (NACTO, n.d.). This can include road width, presence of a bike line and nearness to large scale bike roads (such as the water ways that ring around New York)

References

Citibike. n.d. "Citibike System Data." https://www.citibikenyc.com/system-data.

MTA. n.d. "NYC Transit Subway Entrance and Exit Data." https://data.ny.gov/Transportation/NYC-Transit-Subway-Entrance-And-Exit-Data.

NACTO. n.d. "NACTO Bike Share Station Siting Guide." https://nacto.org/wp-content/uploads/2016/04/