Optimizing Cab Efficiency:

Analysis of 1.5 Billion Taxi Trips in New York City

Aidan Mullan

University of California, Berkeley

May 16, 2019

Introduction

New York City, the most populated city in the United States, has a bustling transportation system including one of the largest subway networks in the U.S., as well as an arsenal of taxi cabs and ride share vehicles. With such a wide variety of public transit options, potential passengers can easily minimize the amount of time they need to wait for a ride by choosing the best option for their location. If a group of people are not near a train station, they have the option of hailing a cab. If no cab is near, they can use a ride-share service to call for a ride.

For a taxi company to remain competitive within this diverse transit system, they need to operate as efficiently as possible. In an ideal case, this amounts to having a vehicle waiting nearby for every person or group needing a ride. It would seem natural, then, for a New York taxi company to dispatch vehicles to densely populated areas such as Midtown Manhattan in order to ensure that most if not all potential passengers get in a taxi. However, this approach neglects two main elements that are crucial to improving the efficiency of a taxi company.

First, if all available cars were sent to a single area, there would be an excess in the supply of cabs which would result in numerous cabs being without a fare. Cabs without a fare, or "idle" cabs, are not earning any money for the company because there are no passengers. These idle cabs would be better located in less populated areas that may not have many cabs readily available to make pickups.

The second issue with the high population allocation strategy is that dispatching cabs to areas in which the demand for taxi fares is high only considers one half of the taxi fare: the pickup. In order to maximize the efficiency of a taxi company as a whole, the other half of a fare, the drop-off, is equally as important. As soon as a cab delivers its passengers to their intended location, it becomes an idle cab that is now available to pick up a new passenger. If there is a potential passenger in the same location as the cab's drop-off, then the cab has little to no idle time and can maximize it's fare earnings. For the taxi company as a whole, optimal operational efficiency is achieved if every cab drop-off occurs in a location in which a passenger is immediately ready to be picked up. Areas in which there are fewer drop-offs occurring than requested pickups would then require additional cabs to meet the unmet demand. Therefore it makes sense to base the taxi company's dispatch strategy on the difference between the demand for cabs and the supply available from delivering fares.

The aim of the present research is twofold. First, we seek to predict areas of New York City in which there is a mismatch between the supply and demand of taxi cabs. Then, using these predictions, we will design an algorithm that creates an optimal dispatch strategy. This strategy would send idle cabs from areas in which supply exceeds demand to meet awaiting passengers from areas in which demand exceeds supply.

Data

The New York City Taxi and Limousine Commission (TLC) has released the data of all taxi fares for the Yellow Cab and Green Cab taxi services dating back to January 2009. These data exceed 1.5 billion fares

in total. For the purpose of this analysis, we will only use data from Yellow Cabs, since they are the most complete. Records for Green Cab fares only began in August 2013, whereas Yellow Cab fares are recorded back to January 2009. The Yellow Cab data alone amounts to over 900 million trips.

The complete data set contains information regarding the location at which the passenger was picked up and dropped off, the time and date of the pickup and drop-off, as well as several indicators of group size, type of fare, cost, and type of payment. Prior to, and including June 2016, all locations were recorded in latitude and longitude pairs. Beginning in July 2016, locations were recorded as the taxi zone ID in which the pickup or drop-off for the fare occurred. For the sake of precision, this analysis will only focus on the data up to June 2016 in order to use latitude and longitude as indicators of location. After restricting the data to this subset, we still have over 750 million taxi trips.

To better understand how taxi usage varies across time, the data were grouped by half-hour time periods, indexed by season of the year, day of the week, and time of day. When grouping by season, the months from December to February were classified as Winter, March to May as Spring, June to August as Summer, and September to November as Fall. Figure 1 provides a heat map of the pickups and drop-offs that occurred on Monday, June 20th, 2016 between 10:00 and 10:30am. Both maps appear to have similar point densities, although there appear to be more drop-offs than pickups in Midtown just south of Central Park. However, from these maps it is difficult to get an objective sense of how the densities of pickups and drop-offs compare spatially.

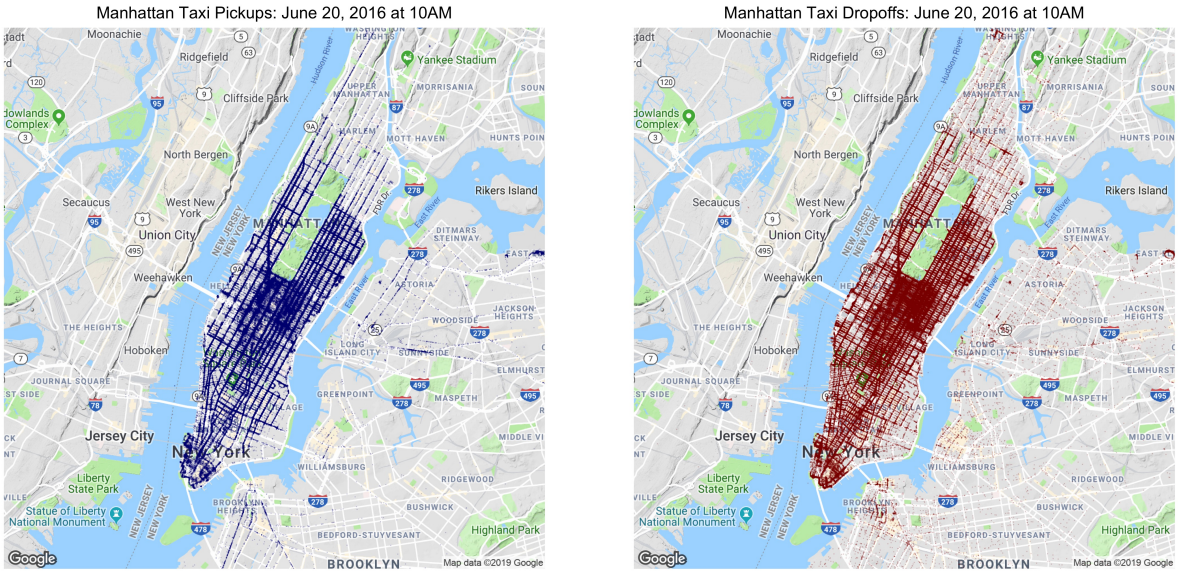

Figure 1: Observed Pickups and Drop-offs for June 20th, 2016 at 10am

The aim of this research is to predict locations in which a mismatch between demand and supply of taxi cabs occurs. To do this, we will use the number of pickups that occurred in a given location as a measure of demand and the number of drop-offs as a measure of supply. The difference between pickups and drop-offs,

then, is a measure of the mismatch we aim to predict.

In order to construct this "net-pickups" density, our data were first grouped into half-hour blocks of time, indexed as described above. This gives us a small window in which to aggregate the number of pickups and drop-offs. Then, we applied a grid system to the data in order to create small regions in which to compare the number of pickups and drop-offs that occurred. The grid was created using a process known as geohashing. Geohashing translates a latitude-longitude pair into a base-32 string that represents a cell in the geohash grid. The length of the string determines the size of the cell and the precision of the grid translation. For this analysis, a geohash string of length 7 was used. This provides grid cells that are 0.0037° by 0.0037°, or approximately 400 square feet in size. In total, this grid provides roughly 10,000 cells that make up New York City. We then count the number of taxi pickups and drop-offs that occurred within a given cell during a given time period. The net-pickups statistic is computed by taking the number of pickups minus the number of dropoffs. For a given cell in the grid, the net-pickups value was computed for every half-hour window during each week from January 2009 to June 2016. This gives us 391 weeks per cell, per half-hour window.

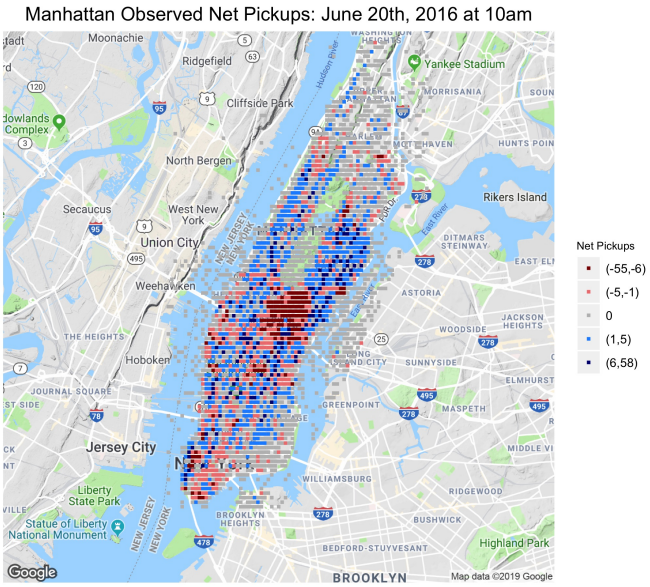

Figure 2: Observed Net-Pickups for June 20th, 2016 at 10am

Figure 2 shows the geohash map for Manhattan on Monday June 20th, 2016 between 10:00 and 10:30am. Here the red squares are cells in which more drop-offs occurred than pickups, indicating an excess in supply. The blue squares are cells in which there were more pickups than drop-offs indicating an excess in demand. Here we can see a cluster of dark red near Central Park in Midtown showing a region in which there is a large excess in supply during this time period. As we move away from Midtown there is a mix of blue and red squares which suggests that neighboring cells may have different signs for the net-pickup values.

Predictive Modeling

To avoid any difference in structure between the number of pickups and drop-offs between two different days of the week, or even different times of the day, models were fit for each half-hour segment in every day of the week separately. The results in 336 time periods on which to fit statistical models.

From the net-pickups map in Figure 2, we see that the data have a general spatial trend, where similar

net-pickup values tend to be grouped nearby to one another. Additionally, there appears to be a substantial degree of variability in the net-pickups values of neighboring cells, indicating the presence of variation from the overall spatial trend. As such, there are two main components to our data that are essential to capture in the model: spatial correlation and temporal variability.

Spatial Smoothing:

In order to account for the overall spatial trend present in the data, we first computed the average net-pickups value for a given cell during a given time block. Let $Y_{i,s,d,t,w}$ be the observed net-pickups for cell i during season s, day of the week d, and half-hour t, from week w in the data set. The cell mean for a single half-hour block is then given as

$$\mu_{i,s,d,t} = \frac{1}{n} \sum_{w=1}^{n} Y_{i,s,d,t,w}$$

where n is the total number of occurrences for the given season, day, and half-hour combination in the data-set.

The last complete week in the data, June 20th to June 27th, 2016, was set aside to be used as a comparison for our model predictions. As such, cell means were computed using all weeks up to this testing week, giving us 390 weeks on which to calculate these cell means.

Using these cell means during a given time block, a spatial smoother was applied to generate a spatial average for each cell. Because the spatial correlation appears to be relatively localized, a K-Nearest Neighbors smoother was applied in order to only consider the closest cells to the target cell. Let $N_k(i)$ indicate the neighborhood of size k for cell i. The KNN spatial average for cell i is then given by

$$\hat{\mu}_{i,s,d,t} = \frac{1}{k} \sum_{j \in N_k(i)} \mu_{i,s,d,t}$$

where k is determined by 10-fold cross validation to minimize the mean squared error between the KNN spatial average and the original cell means.

Figure 3 shows the maps of cell means and spatially smoothed averages for Summer Mondays between 10:00 and 10:30am, which corresponds to the half-hour block depicted in Figure 2. Here we see that the cell means don't appear to differ too much from the observed net-pickups given in Figure 2, although there does appear to be a slightly stronger spatial correlation present. In the map of spatially smoothed means, then, we can clearly see the regions of spatial correlation, from the dark-red area south of Central Park indicating excess supply, to the two strips of light blue on either side of the Park representing regions of excess demand.

Temporal Forecasting:

It is apparent from these maps that although we can capture the general spatial trend present in the data, we lose the variability between neighboring cells found in Figure 2. In order to account for this, we first create a deviation statistic that compares the observed net-pickups from their spatial average. This

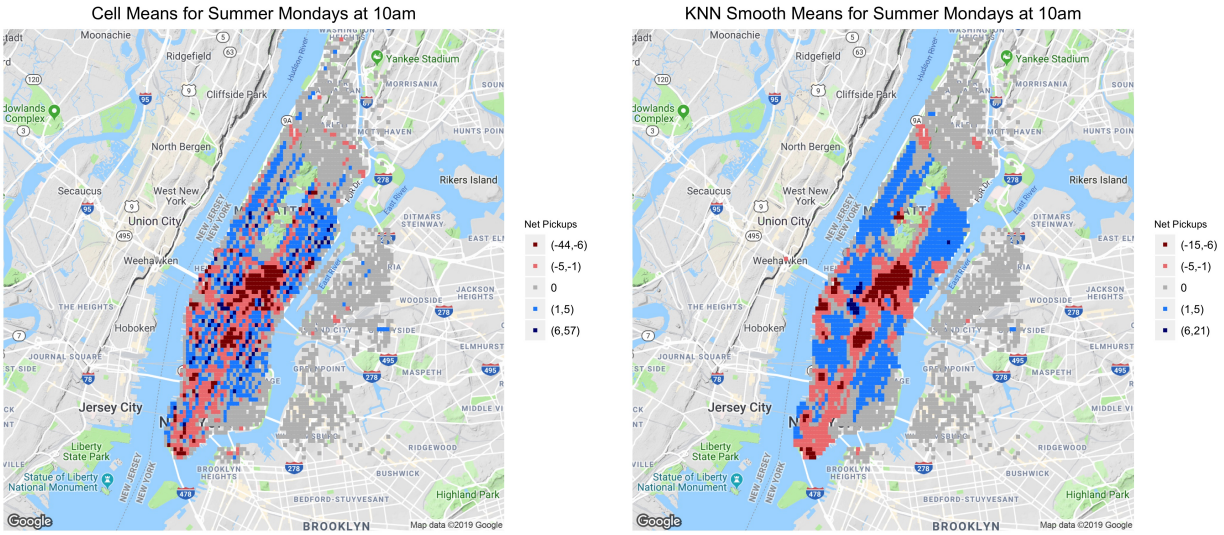

Figure 3: Cell Means and Spatial Averages for Summer Mondays from 10:00 to 10:30am

deviation value is given by

$$\delta_{i,s,d,t,w} = Y_{i,s,d,t,w} - \hat{\mu}_{i,s,d,t}$$

These deviation statistics are then aligned sequentially by subsequent half-hour blocks to create a time series. Here we will denote the Tth entry in the time series for cell i as $\delta_{i,T}$ to avoid indexing issues in which sequential observations fall on different days of the week or seasons of the year. Then, an ARMA model was fit to each cell. EDA of 10 randomly selected cells found that while some cells would be best fit by a lower order model, other cells were better modelled by higher order AR and MA components. To avoid fitting a unique model to each cell, an ARMA(2,2) model was chosen as a middle ground that could fit all cells reasonably well. EDA for selected cells as well as analysis of the ARMA(2,2) fit are given in Appendix A. The ARMA(2,2) model is given by

$$\delta_{i,T} = \phi_{i,1}\delta_{i,T-1} + \phi_{i,2}\delta_{i,T-2} + \theta_{i,1}\varepsilon_{i,T-1} + \theta_{i,2}\varepsilon_{i,T-2} + \varepsilon_{i,T}$$

where $\varepsilon_{i,l} \sim N(0, \sigma_i^2)$ The parameters $\phi_{i,1}$, $\phi_{i,2}$, $\theta_{i,1}$, $\theta_{i,2}$, and σ_i were estimated using maximum likelihood, and a one-step forecast was generated to predict the future deviation statistic. This forecast is computed as

$$\hat{\delta}_{T+1} = \hat{\phi}_1\delta_T + \hat{\phi}_2\delta_{T-1}$$

Then, to retrieve the predicted net-pickups for a given cell at a given time, we add the spatial mean to the predicted deviation:

$$\hat{Y}_{i,s,d,t,w} = \hat{\delta}_{i,s,d,t,w} + \hat{\mu}_{i,s,d,t}$$

Figure 4 provides the observed and predicted net-pickups maps for June 20th, 2016 between 10:00 and 10:30am. Here we find that the predictions are almost indistinguishable from the observed values, indicating

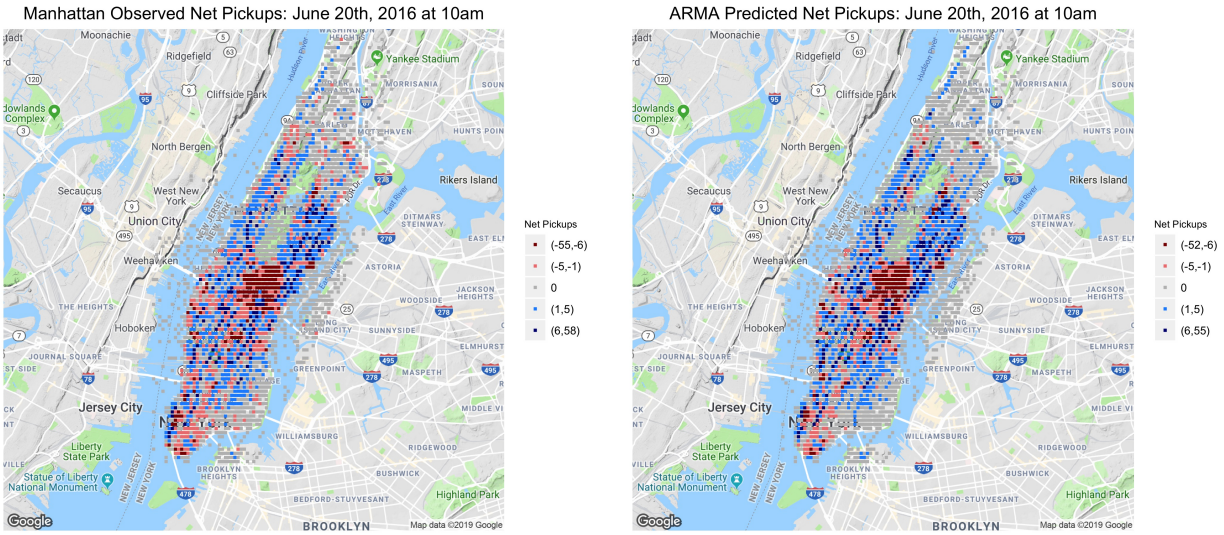

Figure 4: Observed and Predicted Net-Pickups for Monday June 20th, 2016 from 10:00 to 10:30am

a high degree of predictive accuracy. In fact, these predictions have an R^2 of .907 and a root mean squared error of 1.275.

Model Results:

The last complete week in the data, the week of June 20th to June 27th, 2016, was set aside from the rest of the data-set to be used as a test set on which to compare our model predictions. Since all of these days fall in the summer season, we have 336 half-hour blocks of time on which to construct our model. Using the process detailed above, predictions were generated for every cell in each of these 336 blocks of time. R^2 and root mean squared error (RMSE) statistics were computed for each half-hour set of predictions in order to determine the model accuracy. These measures are given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad R^2 = \frac{(\sum_{i=1}^{n}(Y_i - \bar{Y})(\hat{Y}_i - \bar{\hat{Y}}))^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \sum_{i=1}^{n}(\hat{Y}_i - \bar{\hat{Y}})^2}$$

Table 1 provides summary statistics for the accuracy of our model. On average, the model returns an R^2 of .771 , which indicates that the model captures over 75% of the variability present in our data. Additionally, we find the average RMSE to be 2.584, which suggests that on average, our predictions deviate from the observed values by approximately 2.5 net-pickups.

In addition to model accuracy, the overall fit of the model was evaluated through the model residuals. For a given half-hour window, each cell returned a residual computed as the difference between the predicted and observed value. Figure 5 provides a spatial map of the residuals from June 20th, 2016 between 10:00 and 10:30am in order to check for any spatial correlation that went unaccounted in the modeling process.

Table 1: Model Performance Summary

	Minimum	Mean	Maximum
R^2	.536	.771	.952
RMSE	0.946	2.584	4.570

Here we find that there does not appear to be any substantial spatial correlation present, although we do see that larger residuals, represented by darker colored cells, seem to be more frequent in areas that had more extreme net-pickup values. This would suggest that most of the spatial pattern in the data is accounted for by the spatial smoothing portion of our model.

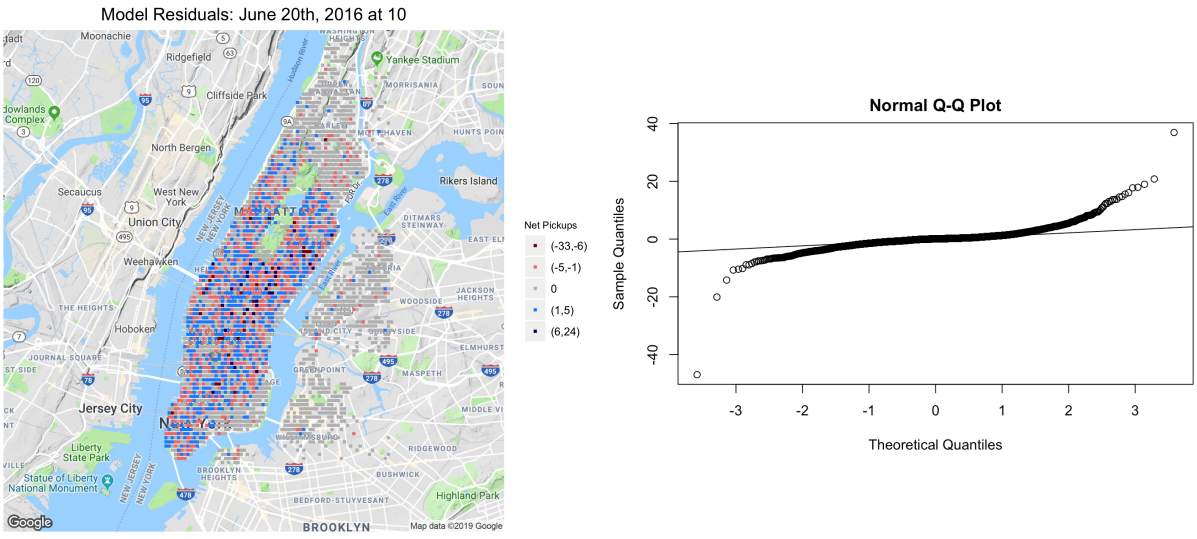

Figure 5: Residual Plots for Monday June 20th, 2016 from 10:00 to 10:30am

In addition to spatial residuals, we also need to check the general distribution of our residuals, given by the quantile-quantile plot in Figure 5. Here we find that the distribution has heavy tails relative to the normal distribution. One possible reason for this is that the error terms in our ARMA model are assumed to be Gaussian white noise, whereas these residuals may indicate that t-distributed errors may be more appropriate. However, both the spatial residuals as well as the residuals from the time series (Appendix A) suggest that our model is a good fit for the data, and the non-normality will not affect the model interpretation.

Dispatch Algorithm

Once predictions are generated, the algorithm to optimally dispatch idle cabs needs to match areas of excess supply, indicated by negative net-pickups, to areas of excess demand, indicated by positive net-pickups. For simplicity, the dispatch algorithm will be described for a given half-hour window of time. In this case,

let $\hat{Y}_i$ be the predicted net-pickups for cell i in this time window.

Let $D = \{i; \hat{Y}_i > 0\}$ be the set of all cells that have a positive net-pickups prediction and $S = \{j; \hat{Y}_j < 0\}$ indicate all cells with a negative net-pickups prediction. D is then the demand region and S is the supply region. For a cell j in the supply region, S_j indicates the number of cabs available for dispatch. Similarly, D_i indicates the number of passengers available for pickups in demand region cell i. These values are computed as

$$S_j = I(j \in S)(-\hat{Y}_j) \qquad\qquad D_i = I(i \in D)\hat{Y}_i$$

Now let d_{ij} be the distance between cell i and cell j, and let x_{ij} be the number of taxis dispatched from cell j to cell i. To minimize computation in this step, we only need to compute the pairwise distances between one cell in the demand region and one cell in the supply region, instead of computing all possible pairwise distances of all cells in the grid. In order to minimize the total distance travelled by all dispatched cabs, the algorithm minimizes the objective function

$$C(x) = \sum_{i \in D} \sum_{j \in S} d_{ij} \cdot x_{ij}$$

However, we impose constraints on this optimization depending on the total excess of demand or supply across all cells. If the total excess in demand is greater than the total excess in supply for a given half-hour window, then we impose the constraints

$$\sum_{i \in D} x_{ij} = S_j \text{ for all } j \in S \qquad\qquad \sum_{j \in S} x_{ij} \leq D_i \text{ for all } i \in D$$

to ensure that all idle cabs are dispatched to a new passenger.

Conversely, if the total excess in demand is less than the toal excess in supply for a given half-hour window, we impose the constraints

$$\sum_{i \in D} x_{ij} \leq S_j \text{ for all } j \in S \qquad\qquad \sum_{j \in S} x_{ij} = D_i \text{ for all } i \in D$$

in order to guarantee that all awaiting passengers are given a ride.

These constraints verify that we are making the maximum possible connections between the idle cabs needing a passenger and the passengers needing a ride. Once we have determined the set of x_{ij} that minimize the total distance traveled by all available cabs, we can construct the dispatch matrix M, where $M_{ij} = x_{ij}$. The entries of this matrix are exactly the optimal number of cabs sent from supply cell j to demand cell i. Moreover, the jth column of M represents the path of dispatch for every available cab in supply cell j, and the ith row of M represents the starting location of every cab being sent to demand cell i.

Principal Component Analysis

Once we have constructed our dispatch matrix M, we can then use principal component analysis to get an idea of the general flow of dispatch. Suppose there are m supply cells and n demand cells. We perform

a singular value decomposition on this matrix to get

$$M_{mxn} = U_{mxm}\Sigma_{mxn}V^T_{nxn}$$

where $\Sigma_{mxn} = \text{diag}(\sigma_i)$. We can then approximate our dispatch matrix M by only considering the first k singular values of Σ, which are given by $\{\sigma_1 \dots \sigma_k\}$. This approximation becomes

$$\hat{M}_{mxn} = U_{mxk}\Sigma_{kxk}V^T_{kxn}$$

To determine the value of k to use for our approximation, we consider the relative size of each eigenvalue in Σ, given by

$$\alpha_p = \frac{\sigma_p^2}{\sum_{i=1}^{n} \sigma_i^2}$$

By ordering the α_p, decreasing in size, and computing their cumulative sum, we get an measure of the additional information provided by each subsequent eigenvalue, which represents the importance of the dispatch flow for a given cell. Typically k is determined by the point at which additional α_p provide diminishing returns to the cumulative sum. The k chosen principal components, then, represent the k cells that control most of the flow of dispatch in our algorithm.

Using the predictions generated for Monday, June 20th, 2016 from 10:00 to 10:30am, we find that the first 80 principal components account for approximately 80% of the dispatch flow provided by our algorithm. The map of our estimated dispatch $\hat{M}$ is given in Figure 6. Here we find that most of the idle taxis come from Midtown and need to be allocated to Upper East Side and Upper West Side. There are also some cells towards the south of New York City in Tribeca and the Financial District that are prominent in our dispatch model as well.

Figure 6: Estimated Dispatch for June 20th, 2016 at 10am

Discussion

This research provides a method by which taxi companies can predict the mismatch between the supply of taxi cabs and the demand for rides, and then better allocate these idle taxis to pick up a new fare. Doing so not only improves the efficiency of a taxi company as a whole, but also reduces the time passengers need to spend waiting to get a ride. This algorithm of optimizing efficiency lends itself to real-time updates in two main ways.

First, the predictions are generated for every half-hour, which is fairly granular. Since the New York City traffic and taxi usage fluctuates even between half-hours, the model can be applied regularly to give frequent updates to the dispatch

strategy being used by the cab company. Predictions are generated every half-hour and a new dispatch strategy is produced. As new data is introduced to the model, the dispatch strategy changes to accommodate the new patterns in taxi usage. As cabs are dispatched to various areas of New York City, inevitably there will be new areas created in which a demand-supply mismatch occurs. The algorithm can then reconfigure the dispatch strategy to account for these new regions of mismatch.

The second major benefit of this model is that the predictive step only relies on the previous hour of taxi usage in order to predict the future half-hour. As such, predictions require fairly little computation to generate, and the model can be updated in real time. If any unusual changes in taxi patterns occur, the model will be sensitive to them and update accordingly. Moreover, these sudden changes will not create lasting impacts to future predictions, assuming that traffic patterns return toward their norm.

Future research may seek to improve upon this algorithm in a variety of ways. No covariates were included in this analysis. Instead, it is implicitly assumed that past observations are sufficient to predict the future. Predictions may become more accurate by including the addition of variables such as weather or holidays into the model. Further, the spatial computation step assumed that every cell had a true net-pickups value that was independent of any temporal trend. Future improvement to this model may consider the cell average as a linear trend when computing spatial correlation between cells.

Appendix A: ARMA(2,2) Model Fit for Select Cells

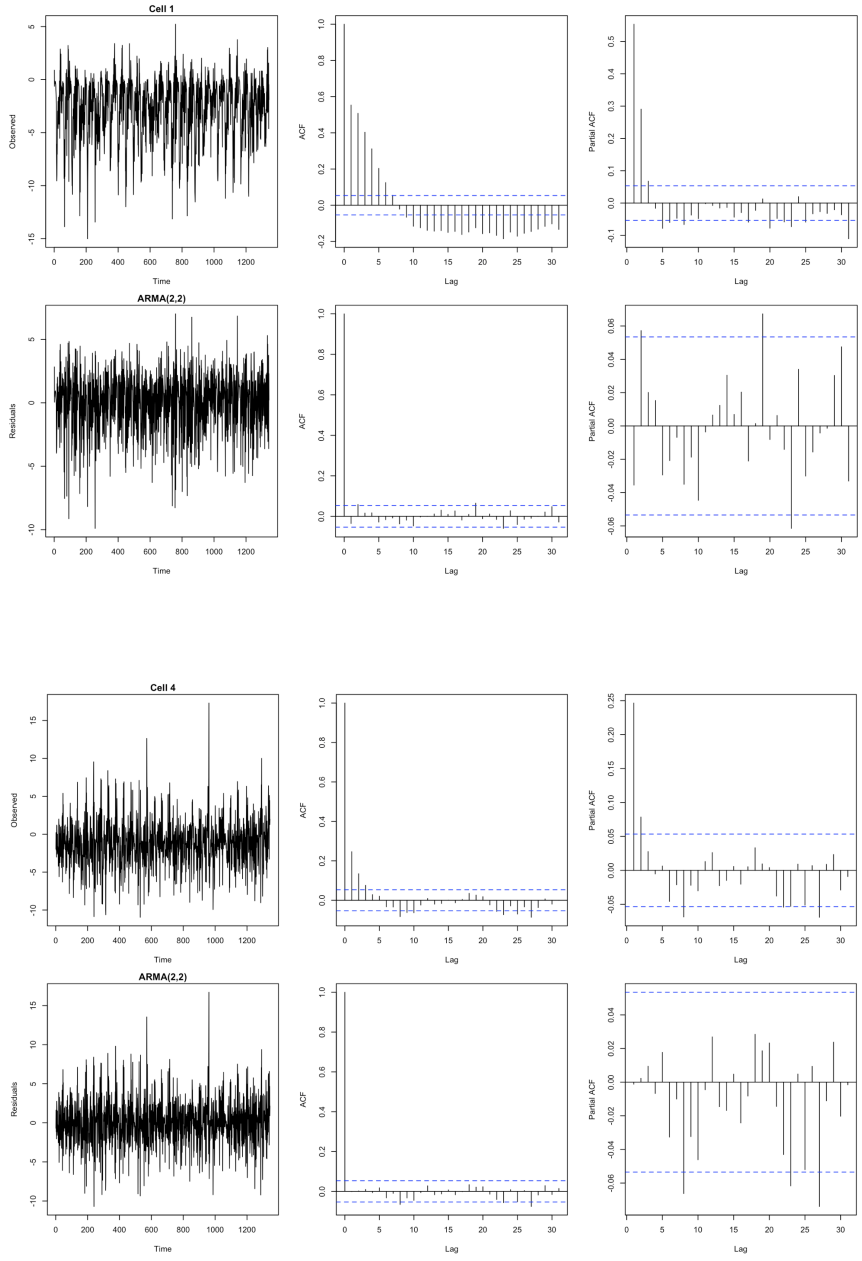

Figure A1: Deviation Series EDA and ARMA(2,2) Model Fit

Appendix B: Data Sources

The raw taxi data used in this research can be found at
https://www1.nyc.gov/site/tlc/about/tlc-trip-record-data.page

The compressed data and the code used for analysis can be found at at my github
https://github.com/mullana2/NYTaxi