Optimizing Efficiency:

Analysis of 1.5 Billion Taxi Trips in New York City

Aidan Mullan

University of California, Berkeley

April 21, 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 to help the numerous citizens get around in their day-to-day life. 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 maximize their operational efficiency within this diverse transit system, they would ideally have 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 for 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 present research seeks to model the difference between taxi demand and supply in New York City by analyzing the locations of cab pickups and drop-offs. These models will then be applied to generate predictions for future "net-pickup" values with the aim of creating a useful tool for improving taxicab allocation and efficiency.

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 750 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.

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.

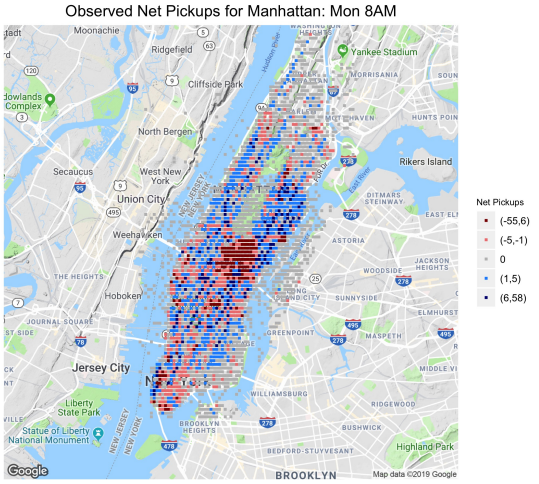

Figure 1: Observed Net-Pickups for June 20th, 2016 at 8am

In order to construct this "net-pickups" density, our data were first divided based on day of the week and half-hour of the day. This gives us a small window of time 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 over 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 390 weeks per cell, per half-hour window.

Figure 1 shows the geohash map for Manhattan on Monday June 20th, 2016 between 8am and 8:30am. 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.

Methods

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. During a given half-hour period, models were fit using the net-pickups value computed for each cell in the grid, every week from January 2009 to June 2016. We fit three types of models to the net-pickups data to determine the best model for generating predictions: k-nearest neighbors regression, ARIMA, and Holt-Winters exponential smoothing.

K-Nearest Neighbors

The New York taxi data are inherently spatial. It would make sense, then, to model the net pickup values using a spatial model. As we have constructed a grid system in order to compute the net pickup values, we used a K-Nearest Neighbors regression model to predict the net-pickups. Here we assume that for a given half-hour window, each cell i has a true net-pickup value that is independent of month or year. Let $Y_{i,t}$ be the observed net pickup value computed in cell i for week t. We can use a sample $\{Y_{i,t}\}_{t=1}^{T}$ to approximate the true value in the cell. We first compute the average value in cell i from week $t = 1$ to $t = T$:

$$\tilde{Y}_i = \frac{1}{T} \sum_{t=1}^{T} Y_{i,t}$$

Next, the k closest cells to cell i determined by geographical distance are grouped into the neighborhood of cell i, denoted by $N(i)$. Distance was computed based on the latitude and longitude values for the center point of each cell, normalized to have mean 0 and variance 1. For the k-nearest neighbors model, the predicted value in cell i for week $T + 1$ is given by:

$$\hat{Y}_i = \frac{1}{k} \sum_{j \in N(i)} \tilde{Y}_j$$

For our data, each cell in a half-hour window has observations for 390 weeks. The last week of June 2016 was set aside to be the test set on which to evaluate our model accuracy, which gives us 389 weeks as our training data ($T = 389$). The value of k to be used in generating predictions for the test set was determined by 10-fold cross validation on the training set, the set of $\tilde{Y}_i$ values, based on the root mean square error of predictions. This choice of k is not necessarily the same for all half-hour time periods.

ARMA Model

In Figure 1, we saw that neighboring cells in the Manhattan grid often had differing signs on the net-pickups statistic. This may suggest that there is little spatial correlation present in our data. For this reason,

it may be the case that each cell can be treated as independent. Then, since we have longitudinal data for each cell, we can model the cells using independent time series models. This amounts to fitting a different time series model for each cell in the grid, for each of our 336 half-hour segments. Our first approach is to use ARMA models. To do this, we first assume that the net-pickups statistic for each cell is weakly stationary with a constant mean and variance:

$$\mathbb{E}[Y_{i,t}] = \mu_i \qquad\qquad \mathbb{V}[Y_{i,t}] = \sigma_i^2$$

Then, each cell is fit to its own ARMA model. During the model fitting process, AR and MA components from 0 to 5 were considered. The best model for a given cell was determined by a comparison of Akaike Information Criteria (AIC). For a given half-hour, let $Y_{i,t}$ be the net-pickups for cell i during week t. The ARMA model is given by

$$Y_{i,t} = \sum_{k=1}^{5} \phi_k Y_{i,t-k} + \sum_{k=1}^{5} \theta_k \epsilon_{i,t-k} + \epsilon_{i,t}$$

where $\epsilon_{i,j}$ is a white noise term with mean 0 and variance ξ_i^2. For the purpose of selecting an optimal model, the estimates $\hat{\phi}_k$ and $\hat{\theta}_k$ are allowed to be 0. However, if $\hat{\phi}_k = 0$, then $\hat{\phi}_{k+j} = 0$ for all j. The same applies if $\hat{\theta}_k = 0$. This ensures that all $Y_{i,t-k}$ and $\epsilon_{i,t-k}$ values included in the model are sequential. Since the epsilon terms are all mean 0, the forecast for $Y_{i,T+1}$ is given by

$$\hat{Y}_{i,T+1} = \sum_{k=1}^{5} \hat{\phi}_k Y_{i,T-k+1}$$

As with the k-nearest-neighbors model, the last week of June 2016 was used to test model accuracy. The models were fit using the net-pickup values up to but not including the last week in June ($T = 389$). The ARMA model for each cell was chosen based on AIC, and the forecast value was computed. The ARMA(p,q) model used was not necessarily the same for each cell, or for the same cell in different half-hour windows.

Holt-Winters Exponential Smoothing

One of the main assumptions with ARMA models is that the series being modelled is weakly stationary. Since we have thousands of series for each of 336 time periods, we cannot confirm that this assumption always holds. To accommodate for this, we also used a Holt-Winters Exponential Smoothing model to generate predictions for each cell. Exponential smoothing methods perform better on non-stationary processes, and may provide more accurate predictions for the net-pickups statistic in the case that most of the cells have non-stationary processes underlying their time series. Additionally, the exponential smoothing process uses information from all past observations in generating predictions, whereas ARMA processes only require the past k values.

For a given cell i, the Holt-Winters exponential smoothing model is given by

$$Y_{i,t+1} = l_t + b_t$$

$$l_t = \alpha Y_{i,t} + (1-\alpha)(l_{t-1} + b_{t-1})$$

$$b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$$

where l_t estimates the level of the time series at time t and b_t estimates the slope. $\alpha \in [0,1]$ is a smoothing parameter for the level of the series and $\beta \in [0,1]$ is a smoother for the overall trend. Here the level estimate is a weighted average of the observed Y_t and the one-step ahead forecast for Y_t computed on l_{t-1} and b_{t-1}. We can clearly see in this formula that exponential smoothing is inherently recursive, utilizing information from all past observations with decaying weight as the lag increases.

The last week of June 2016 was set aside when computing the parameter estimates $\hat{\alpha}$ and $\hat{\beta}$ for the model. These estimates were computed by minimizing the root mean squared error of the series predictions from time $t = 1$ to $t = T = 389$. The forecast for time $T + 1$ was then computed as

$$\hat{Y}_{i,T+1} = \hat{l}_T + \hat{b}_T$$

where $\hat{l}_T$ and $\hat{b}_T$ are calculated using $\hat{\alpha}$ and $\hat{\beta}$.

Results

The three statistical methods were run using all New York grid cells for each of the 336 half-hour time windows in our data. For a given time period, the performance of each prediction method was evaluated using the root mean squared error (RMSE) and R^2 values. Let Y_i be the observed net-pickups for cell i during the last week of June 2016, and let $\hat{Y}_i$ be the predicted net-pickups for the same time. We compute the RMSE and R^2 values as

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

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

where $\bar{Y}$ is the average of the observed net-pickups and $\bar{\hat{Y}}$ is the average of the predicted net-pickups. For our purposes, the R^2 statistic tells us how much of the variability in the observed data we capture in our predictions, whereas the RMSE gives us an estimation of how far our predictions deviate from the observed net-pickups.

Summary statistics for the performance of the three models is given in Table 1. From this table we find that k-nearest neighbor modeling returned an average R^2 of 0.386 with a range from 0.080 to 0.747. This tells us that on average, our spatial modeling was only able to capture approximately one-third of the variability in the observed data.

Table 1: Model Performance Summary

	Minimum	Mean	Maximum
KNN R^2	.080	.386	.747
ARMA R^2	.553	.764	.950
Smoothing R^2	.536	.771	.952
KNN RMSE	0.655	2.703	7.033
ARMA RMSE	0.446	1.299	2.123
Smoothing RMSE	0.433	1.271	2.056

Comparatively, the ARMA predictions returned an average R^2 of 0.764 and the Holt-Winters smoothing predictions returned an average R^2 of 0.771. Both of these methods perform drastically better than the KNN predictions. Moreover, both of these time series methods generated more accuracte predictions on average, as detailed by the lower mean squared error. When we look at the two time series methods, the exponential smoothing predictions performed marginally better than the ARIMA forecasting in both R^2 and root-mean squared error. This would suggest that the grid cells are better modelled as non-stationary series. On average, both time series models were able to capture over 75% of the variability in the net-pickups, and had predictions that were off by just over 1 net-pickup.

From the R^2 and RMSE statistics, we can see that the k-nearest neighbors method was out-performed by both of the time series methods. When we look at the predicted net-pickup values as a spatial distribution, it becomes clear why this is the case. Figure 2 gives the KNN predictions of the net-pickups for Monday June 20th, 2016 between 8am and 8:30am. For this time window, k-nearest neighbors returns an R^2 of .503 and an RMSE of 2.902.

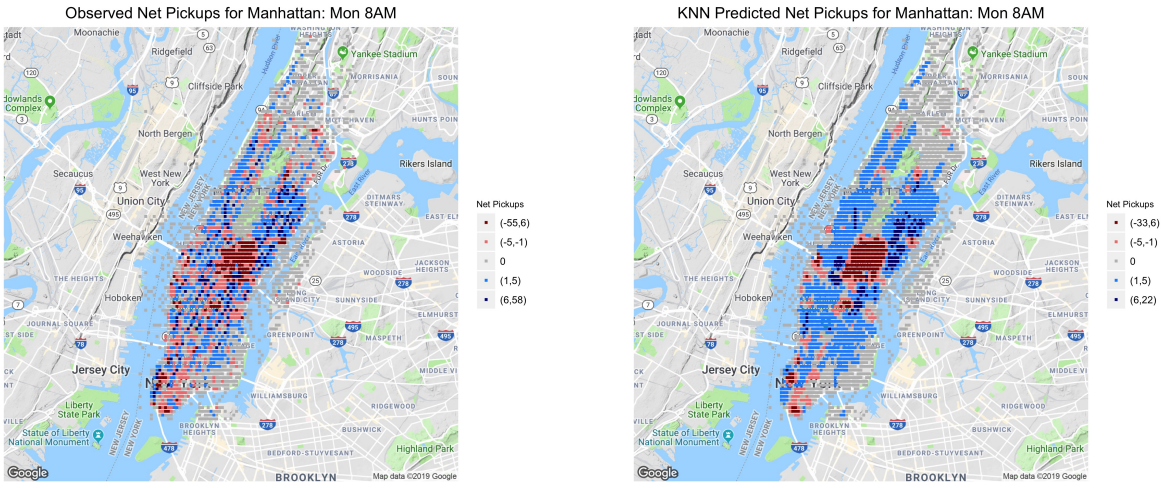

Figure 2: Observed and KNN Predicted Net-Pickups for June 20th, 2016 at 8am

Overall, the KNN predictions seem to capture the general spatial trends in the data. In Figure 2, we see the large dark red portion around Midtown and Central Park in both the observed and predicted maps, with areas of light blue in the surrounding region. As we move further from central Manhattan, we find more gray cells. However, k-nearest neighbors by nature smooths our data by generating predictions from the average of neighboring cells. We can clearly see this in the prediction map, where there are large areas of a solid color indicating that all neighboring cells are predicted to have similar net-pickups. It is evident in the observed map, however, that the actual net-pickups can vary heavily between neighboring cells.

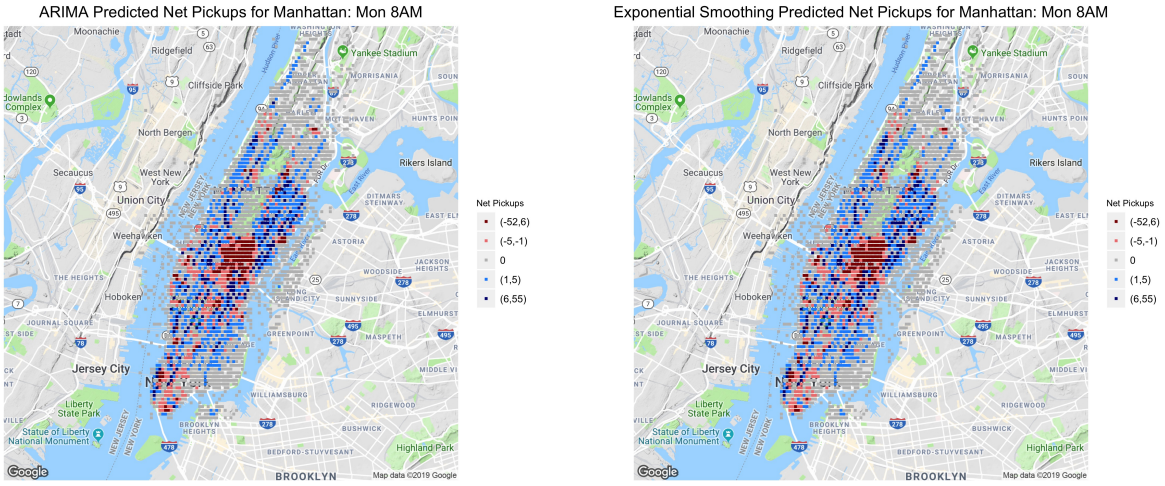

Figure 3: ARMA and Holt-Winters Predicted Net-Pickups for June 20th, 2016 at 8am

When we look at maps of the ARMA and exponential smoothing predictions given in Figure 3, there doesn't appear to much difference between the two models. In fact, we find an R^2 of 0.904 and 0.911 for the ARMA and smoothing predictions, respectively. Both methods capture the overall trend of the data, but also allow for neighboring cells to differ in their net-pickup values.

Residual Analysis

To judge the overall fit of our model, we can look at the residuals for our predictions. Since each half-hour window will produce its own set of model residuals, we will use the residuals from Monday at 8am as an example case.

Here we find a clear difference between the spatial and time series models in the residuals. Figure 4 shows the model residuals plotted against the observed values during the Monday at 8am window for which we created the previous maps.

Here we can see that there doesn't appear to be any discernible pattern in the residuals for either of the time series methods. However, the k-nearest neighbors residuals appear to have some positive linear trend, which suggests that the KNN model does not handle extreme values well. This would make sense, because

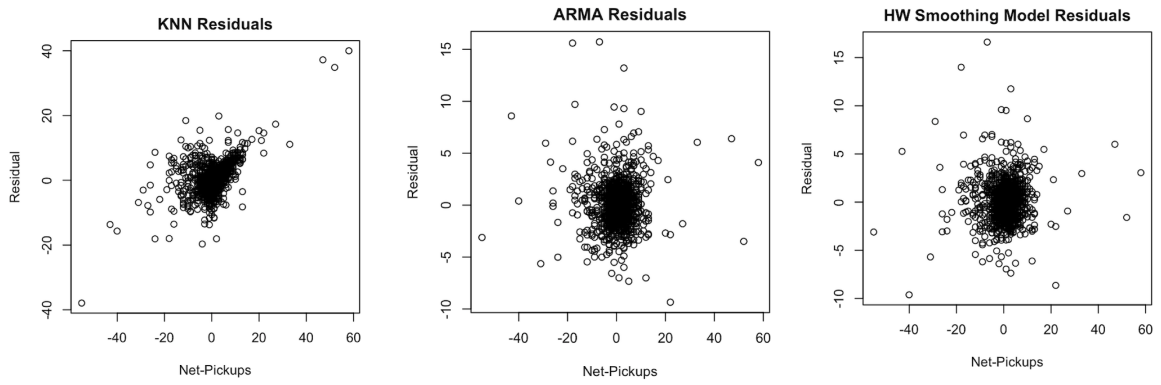

Figure 4: Model Residuals for June 20th, 2016 at 8am

the KNN method relies on averaging to generate predictions, and it is difficult to get extreme values when taking an average. The ARMA and HW smoothing residuals appear to be normaly distributed about 0 with no real association to the observed values. We do find several large residuals across all three plots, which would suggest that there are cells in the New York grid for which our predictions differ greatly from the true net-pickup value. However, the residuals for the KNN model are much larger than those from either of the time series methods, which serves as another indication that our time series predictions are better than the spatial KNN model.

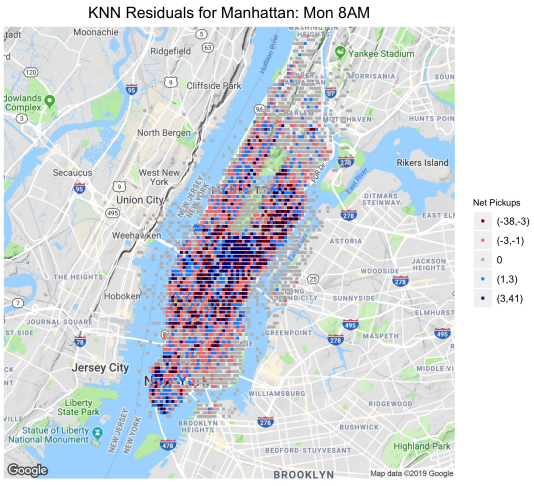

Figure 5: KNN Residuals Map for June 20th, 2016 at 8am

Since the nature of our data are spatial, it is also important to look at the residuals in terms of their spatial orientation relative to one another. Figure 5 provides the residual map for the k-nearest neighbors regression model during the Monday at 8am window, and Figure 6 shows the residual map for the ARMA and Holt-Winters methods during the same time window.

In the KNN residuals (Figure 5), we see a clustering of extreme values toward Midtown and Central Park, where the observed map showed a grouping of large negative net-pickup values. We also find many dark colored squares indicating the presence of extreme residuals in the area surrounding Midtown. This would suggest that there is a structural issue with the KNN regression model, most prominently that it does not handle regions with extreme net-pickup values well.

The time series residuals, mapped in Figure 6, do not have any major noticeable pattern. There appear to be more non-gray cells around Midtown, which may suggest that the time series methods were not able

to predict regions with extreme net-pickup values well, similar to the KNN regression. However, the range of residuals for both the ARMA and exponential smoothing predictions is much smaller than KNN, again indicating that these models are a better fit to the data.

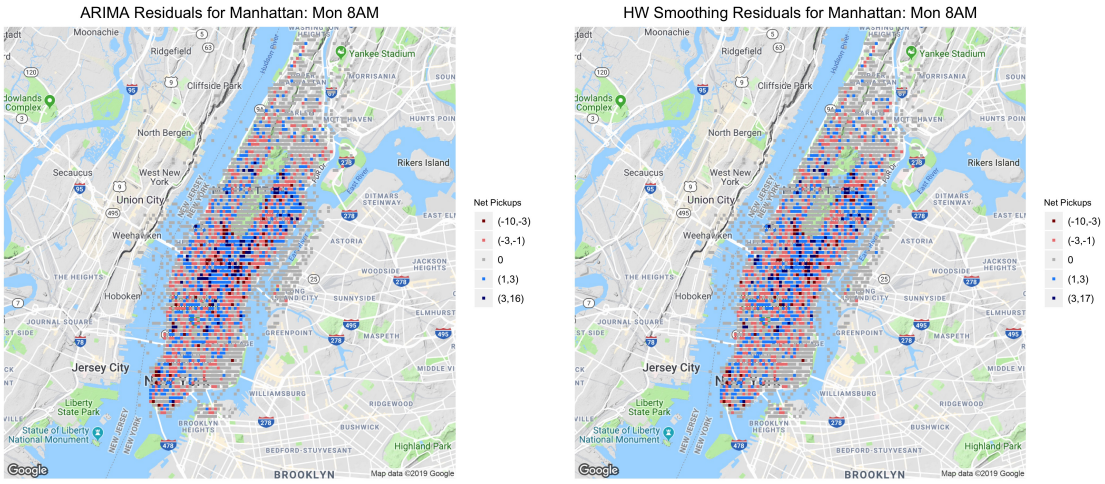

Figure 6: Residual Maps for ARMA and Holt-Winters Models for June 20th, 2016 at 8am

Model Performance by Time

Predictions were generated for each day of the week and half-hour within each day. It is crucial in analyzing model performance, then, to determine if either day of the week or half-hour of the day had any correlation to the accuracy of the models. Figure 7 provides the R^2 and root mean squared error of all three models averaged by day of the week. Figure 8 provides the same summaries, but averaged by half-hour of the day. Looking at accuracy by day of the week, all three models had the highest R^2 values during the middle of the week, on Wednesday and Thursday, and had the lowest R^2 on the weekend. For the time series methods, there does not appear to be any pattern between RMSE and day of the week. However, we find that the KNN regresion had its highest RMSE, and therefore furthest predictions from the true value, on Wednesday and Thursday.

The findings may suggest that there are more regular patterns in cab usage during the week, which would result in more consistent net-pickup values for each week and better predictive accuracy. The weekends, on the other hand, may have a higher variance in net-pickups from week to week, which would make generating predictions using our methods more difficult.

On the whole, the average R^2 and RMSE for both the ARMA and Holt-Winters predictions are relatively constant when averaged across each half-hour window within a given day (Figure 8). We do find an increase in R^2 between 6:00am and 9:30am, which corresponds to the time in which most people are heading into New York City for work. This would suggest that our models perform well with regular patterns of high taxi

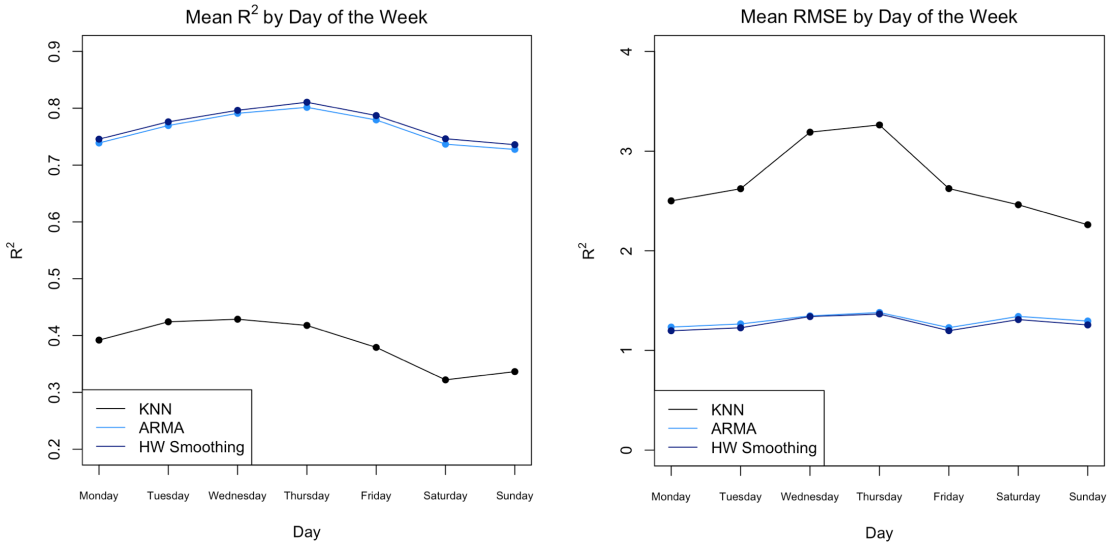

Figure 7: Average R^2 and RMSE by Day of the Week

usage, which makes sense given the nature of time series predictions. It is curious, however, that we do not find a similar increase in accuracy for the end of the work day when many people would be heading home. There is a slight increase in RMSE for both time series models during this period of morning commute, but it does not differ heavily from the other half-hour windows.

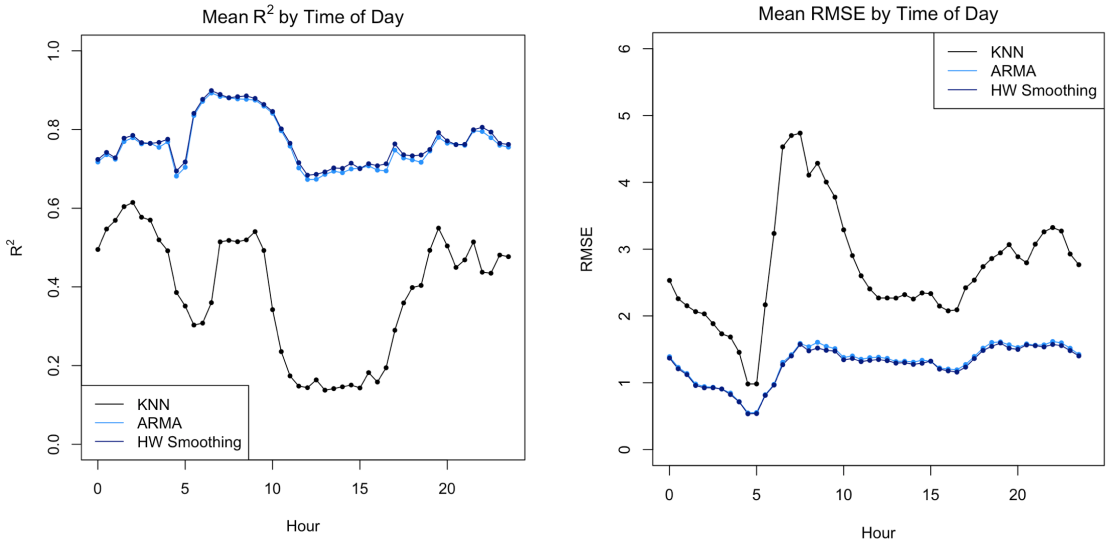

Figure 8: Average R^2 and RMSE by Hour of the Day

The KNN predictions, on the other hand, have a significant decrease in R^2 between 10:30am and 4:30pm. It is unclear from the data why this drop in accuracy occurs. One possibility is that this may be a time in which there is a high spatial variance in cab usage due to people going to specific parts of New York City for lunch. When neighboring cells have largely differing net-pickup values, the KNN performs notably worse

due to the spatial smoothing that generates predictions.

Discussion

The time series predictive methods achieved a relatively high degree of accuracy, with the Holt-Winters exponential smoothing model producing R^2 values upwards of 0.952, which occurred for our model of Tuesdays from 7:00 to 7:30am. This seems to suggest that we are able to predict the future mismatch between the demand and supply for taxi cabs fairly well.

By using our predictive model of mismatched demand, taxi companies can have a better idea of where unaccounted pickups will be, in which case they would be able to send idle cabs to these locations in order to maximize the number of fares being driven and minimize the idle time between the cab's passengers. Moreover, this optimization would minimize the amount of time in which passengers need to wait until they can find a ride.

Since our analysis only utilized the taxis driven by Yellow Cab in New York City, our predictive model also opens the door for competition. By modeling regions in which Yellow Cab are not able to pick up passengers, other taxi companies or ride share services would be able to strategically position idle vehicles to steal potential passengers away from their competition. Although this would harm the efficiency and profits of the Yellow Cab company specifically, the increased competition would greatly benefit the passengers, whose waiting time for a ride would be reduced as long as they do not care which specific taxi company or ride share service they use.

One interesting aspect of applying this predictive model is that the areas in which demand and supply differ will change over time. A taxi company such as Yellow Cab has a limited number of vehicles in the fleet. As such, if the company begins to send cars to areas in which there is excess demand, this will create new areas in which demand is not being met. In this case, the predictive model would need to be updated relatively frequently in order to account for the change in demand and supply. This process would lend itself naturally to time series methods and in particular exponential smoothing, in which the most recent events are weighted the heaviest when forecasting into the future.

It should be noted however, that there are several strong assumptions that are being made with these time series models. Both the ARMA and exponential smoothing models assume that past observations are sufficient in predicting the future. Although we do not need to assume stationarity for the Holt-Winters smoothing model, we still require that no structural breaks or deviations occur in the process generating our time series. Additionally, there are no covariates present in the model. Our forecasts from the time series models are in essence just linear combinations of past events. It may be the case that weather patterns, holidays, or special events cause the usage of taxi cabs or public transportation to change, which would not currently be captured by our models.

Moreover, we found that the time series models had a drop in predictive accuracy during the weekends, which may be caused by increased variance in week-to-week taxi usage. Since the ARMA and exponential

smoothing models assume that past events can be used to predict the future, our forecasts are unable to capture this variance. Instead we can forecast the trend in net-pickups, which will usually differ from the observed value more heavily when the variance is high.

Future research into this data may look to add covariates into the analysis to try and account for additional variability in the data. Additionally, our analysis restricted the data to Yellow Cabs only. The New York City TLC also published data for the Green Cab taxi company beginning in August 2013, as well as select ride-share data beginning in 2016. Incorporating these additional services may provide more accurate and up-to-date predictions, since the public transit system has most likely changed over the past few years.