Predicting Taxi Pickups in Cities: Which Data Sources Should We Use?

Austin W. Smith
Analytics & Data Science
University of New Hampshire
Durham, NH 03824, USA
aws1016@wildcats.unh.edu

Andrew L. Kun
ECE Department
University of New Hampshire
Durham, NH 03824, USA
Andrew.Kun@unh.edu

John Krumm
Microsoft Research AI
Redmond, WA 98052, USA
jckrumm@microsoft.com

UbiComp/ISWC '17 Adjunct, September 11-15, 2017, Hawaii, USA.

ACM 978-1-4503-3575-1/15/09...$15.00.
http://dx.doi.org/xx.xxxx/xxxxxxx.xxxxxxx

Abstract
In this paper we use machine learning to evaluate the predictability of taxi pickups in New York City's busiest neighborhoods using data available from the Taxi and Limousine Commission and from Twitter. We found that these pickups can be accurately forecast, and we show which features work well for forecasts, including geotagged Twitter posts in some cases.

Author Keywords
Urban Mobility; Intelligent Transportation; Machine Learning; Urban Sensing; Pervasive Urban Computing; Transportation Demand Forecasting

ACM Classification Keywords
J.4 Social and Behavior Sciences.

Introduction
In many urban areas the public can take advantage of a number of public transportation options, from buses, to subways, to ferries, to taxis and ride-sharing services. Among these options, taxis and ride-sharing services play an important role, because unlike the other options, they are flexible both in location and in time. That is, a person who wants to ride a taxi or ride-sharing vehicle can (ideally) ask for a pickup at any location in a city, and at any time of the day.

Yet, this location and time flexibility will only be possible if the companies that provide the transportation services can accurately predict the number of pickups in different locations and at different times. Such accurate predictions would allow the companies to field a sufficient number of vehicles at the appropriate places and times in order to handle the demand by the traveling public. It is possible that human taxi drivers can intuitively predict demand by location and time, based on their prior experiences. However, we suspect that it is possible to design predictive algorithms that can outperform this human intuition. Furthermore, such algorithms will be indispensable in the (perhaps not so distant) future when vehicles are automated, and a system controlling the vehicles needs to predict demand [1]. This paper explores the question: what are the data sources that can help us to accurately predict taxi pickup demand in a city, both for different locations, and for different times of the day? Here, we attempt to answer this question by exploring data that is available for New York City. Specifically, our hypothesis was that the demand for taxi pickups for a given place and time in New York City can be predicted from the number of (a) pickups, and (b) tweets at that place over a relatively short time period preceding the time of interest.

Related work

The data made publicly available by the New York City (NYC) Taxi and Limousine Commission (TLC) [2] has been the subject of a number of studies. Primarily this research has focused on descriptive aspects of the data, such as looking at where riders are travelling or how much they are paying, along with other statistics and visualizations. In early releases of the data, a unique identifier was provided for individual drivers, allowing for the analysis of specific drivers, which yielded interesting findings. For example, using these identifiers, it is possible to determine which drivers are the fastest or most efficient. However, the unique identifier was later stripped from the data. Other studies used the data to evaluate patterns of human movement, including the impact of large storms on cab rides.

Similarly, social media is very frequently the subject of research. Often the content is what is considered, rather than the counts of occurrences. The use of Twitter to evaluate human activity in Manhattan has been thoroughly studied and documented.

Ferreira et al. used the NYC taxi trip data as a sensor for activity within the city [3]. Their main focus was on developing a system for querying and visualizing this data that is more in-depth and useful than standard analytics queries. Due to the size of the data, they developed a system of storage, querying and visualization allowing for interactive plotting of various visualizations. The system they created can generate heatmaps, plots and other investigative visualizations. A particular example they showed was a comparative heatmap showing the impacts of Hurricanes Sandy and Irene on cab ridership in lower Manhattan.

Patel and Chandan focused on the technical solutions needed to work with the NYC taxi data [4]. Since 2014 there were around 180 million taxi rides, and the authors saw the need for the use of big data tools to analyze the data in a reasonable amount of time. Using technologies such as Hadoop, MapReduce, Hive and Pig, the authors answered questions such as which driver travelled the most distance, which driver

Vendor ID
Pickup Date Time
Drop-off Date Time
Pass. Count
Trip Distance
Pickup Longitude
Pickup Latitude
Rate Code ID
Store and Forward Flag
Drop-off Longitude
Drop-off Latitude
Payment Type
Fare Amount
Extra
MTA Tax
Improve. Surcharge
Tip Amount
Tolls Amount
Total Amount

Table 1 These are the fields available for each trip in the NYC taxi data.

collected the most fares, and what region has the most drop-offs. They also looked at various descriptive plots and heatmaps. The authors were mostly concerned with evaluating the technical aspects of working with such a large amount of data efficiently.

Krumm and colleagues explored how Twitter activity can be used to classify regions of the city into land-use profiles [5]. Furthermore, they explored how Twitter activity in one geographic area of Manhattan might predict activity in another area. In this work, they used the number of tweets in cells of a grid placed over the NYC metro area. They did not use the content of the tweets.

This paper contributes to existing work by building models to predict taxi pickups based on date/time, previous pickup counts, and tweet counts.

Data

In order to successfully predict taxi pickups, we first cleaned, and manipulated two datasets: the NYC TLC dataset, and geotagged tweets from New York City.

Data Cleaning and Preparation

The main goals of the data cleaning and preparation process were as follows:

- Clean data by e.g. remove rows with missing values
- Classify tweets and cab pickups into separate neighborhoods in New York City
- Create a structured data table to allow us to complete the analysis
- Extract data from the database and populate table

Data Cleaning and Preparation: Taxi Data

The first step was to download the taxi data from the TLC website. It came in the form of comma separated value (CSV) files for each month. They provide data for both Green Cabs and Yellow Cabs. Yellow Cabs are the taxis which only pickup in the Manhattan neighborhoods, while Green Cabs pick up passengers in all the outer boroughs. Yellow Cab data is available for the 2009 – 2016 time period, while Green Cab data is available for the mid-2013 – 2016 time period. The entire data set is spread across 113 CSV files. Each of these files contains the fields listed in Table 1. In the work presented in this paper we used both Yellow Cab and Green Cab data spanning five months in mid-2015.

One challenge with the data was that the field names varied slightly from year to year and between the two types of cabs. Additionally, upon initial inspection of the CSV files, there were fields with odd values for both trip mileage and for some of the coordinates. Many trips were found to either originate in the ocean or terminate in the ocean. This was likely due to some sort of anomaly in the GPS sensor in the data logging devices.

In preparation for modeling pickups in neighborhoods, a GeoJSON file sourced from CivicDashboards [6] containing the boundaries of all 428 New York City neighborhoods was used to add in the following features into the data.

- Pickup Neighborhood
- Drop-off Neighborhood
- Pickup Borough
- Drop-off Borough

Neighborhood	Mean Tweets	Mean Pickups	% Pickups
Midtown	140.16	2927.54	0.16
Upper East Side	31.25	2263.43	0.28
Chelsea	67.23	1542.23	0.37
Upper West Side	37.56	1441.51	0.45
Hell's Kitchen	44.01	871.71	0.50
East Village	27.82	656.22	0.53
Theater District	43.46	616.59	0.57
West Village	23.39	615.74	0.60
Murray Hill	6.21	438.83	0.62
LaGuardia Airport	4.23	409.15	0.65
East Harlem	12.37	357.43	0.67
Harlem	26.14	351.91	0.69
Gramercy	9.80	350.52	0.70
Greenwich Village	15.27	348.60	0.72
SoHo	28.92	343.43	0.74
Kips Bay	5.25	337.60	0.76
Financial District	39.62	316.48	0.78
Tribeca	17.98	290.57	0.79
Williamsburg	32.32	289.92	0.81
Flatiron District	11.07	266.89	0.82
Lower East Side	16.68	265.35	0.84
Central Park	21.96	264.52	0.85
Battery Park City	5.61	167.22	0.86
Astoria	11.70	147.78	0.87
Morningside Heights	6.91	141.23	0.88
Long Island City	13.33	136.17	0.89
NoHo	5.60	126.01	0.89
Elmhurst	7.85	101.16	0.90
Bedford-Stuyvesant	18.48	92.65	0.90

Table 2 We used 29 neighborhoods in our analysis, which comprised about 90% of the total taxi pickups in the NYC area.

The process of adding these variables to the data was completed as the files were being loaded into a SQLite database. A script was used to strip rows which had values that were either missing or with trip distances that were extremely large. Once the geocoding process was completed, any rows which were outside the bounds of any neighborhood were also removed. After cleaning the data, it was deposited into the SQLite database with a table for each CSV file.

This process involved a large amount of manual exploration of the data and determining which data fields were common across all the files and what they should be named to ensure continuity between the SQLite tables. Not using any big data technologies and running a simple script on laptops, the runtime of this script was around 15 hours, a number which could surely be improved upon. The final output was a SQLite database which is about 200 gigabytes in size.

Data Cleaning and Preparation: Twitter Data

The Twitter data was queried from the Twitter API for all geotagged tweets, which comprise approximately 1% of all tweets. This data was from the same 5 month date range and area over the city as we used for taxi data, as shown in Figure 1. The data consisted of a timestamp and location for each tweet resulting in about 4 million data points. The same geocoding steps were taken with this data as with the TLC data. We removed all tweets which were not in a NYC neighborhood. Additionally, the timecode was in UTC time, so it was converted to eastern standard time.

Predicting Taxi Pickups

To predict taxi pickups, the data for each neighborhood was manipulated into individual, hour-long instances that included the actual number of pickups in the neighborhood along with several possible predictive features. In total, there were over 60 million instances from all 29 neighborhoods. The specific fields for each instance were:

- Year, Month, Day, Weekday, Hour
- Neighborhood
- Pickup Count
- Tweet Count
- Mean Neighborhood Pickups
- Mean Neighborhood Tweets
- Lag 1, 2, 3, 4, 5 Pickups
- Lag 1, 2, 3, 4, 5 Tweets

Final Modeling Process

The objective for the final model is to predict the "Pickup Count" variable and to determine if a reliable model can be made to predict these factors. Since there are some neighborhoods which have a very low volume of pickups and tweets, only the neighborhoods that comprise 90% of the pickups were considered. Table 2 lists these 29 neighborhoods along with the mean hourly pickup count, and Figure 1 shows in red the subset of 29 neighborhoods that we focused on for the analysis.

Predictors	Mean MAE [pickups per hour]
Time & Pickups	44.76
All Variables	45.70
Cab Pickups Only	59.87
Time & Tweets	63.55
Time Only	63.92
Tweets Only	185.57

Table 3. These are the overall mean absolute prediction errors (MAE) over all the 29 neighborhoods using different sets of features.

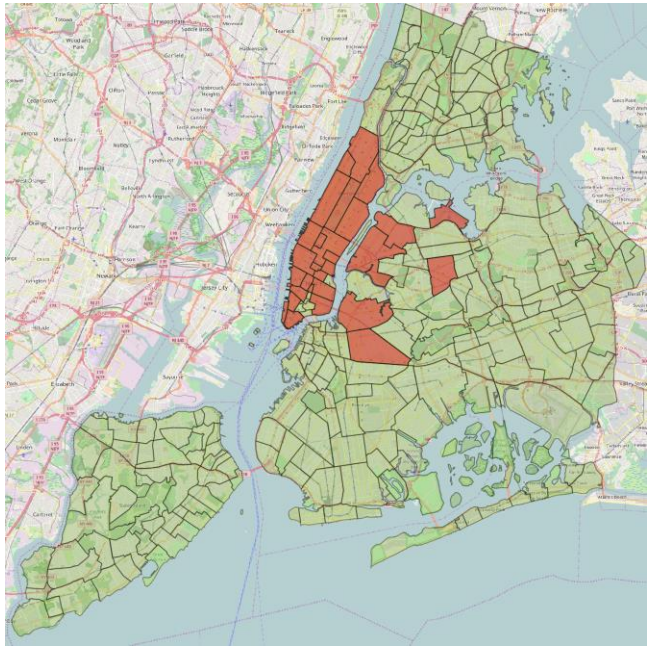

Figure 1. The 29 neighborhoods used in our analysis are shown in red. The green neighborhoods show the other areas covered by the NYC taxi data.

We elected to run a series of models to determine which set of predictors would yield the best results. The metric we chose to use to evaluate the quality of the models is Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Here y_i is the actual pickup count, $\hat{y}_i$ is the estimated pickup count, and n is the number of estimates.

For these models, we trained a random forest regression model on several different subsets of our available features in an effort to understand which features work best.

The feature sets we used are as follows:

- *Cab Pickups Only:* The 5 taxicab lag variables
- *Tweets Only*: The 5 tweet count lag variables
- *Time Only*: Time variables (hour, weekday etc.)
- *Time & Pickups*: Time and taxicab lag variables
- *Time & Tweets*: Time and tweet lag variables
- *All Variables*: All possible variables

We created a model using each of these combinations of predictors for each neighborhood using k-fold cross validation with 10 folds total.

Results

Table 3 shows the MAE for all 29 neighborhoods. Each row in the table represents a different group of predictors. It shows that "Time & Pickups" was the best combination of features, a narrow improvement over using all the variables. In general, Table 3 shows that it is possible to predict the number of taxi pickups with a mean absolute error of less than 50 pickups per hour over all neighborhoods.

Figure 2 summarizes the results of the tests over all the neighborhoods we tested. In general, "Time & Pickups" works best along with using all the available variables. "Time Only" works about as well as "Time & Tweets", while "Tweets Alone" is the worst.

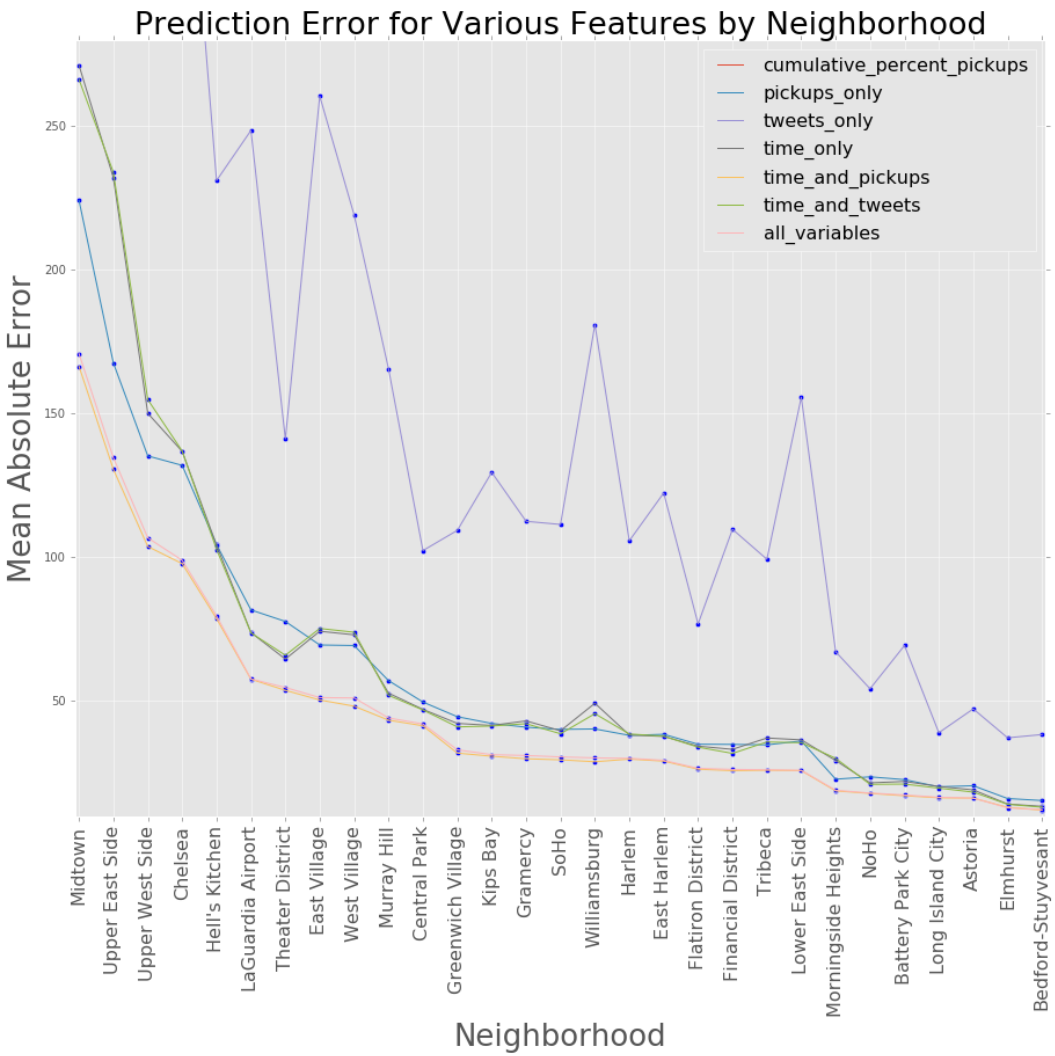

Figure 2. The mean absolute error (MAE) of prediction varies from neighborhood to neighborhood depending on the prediction features.

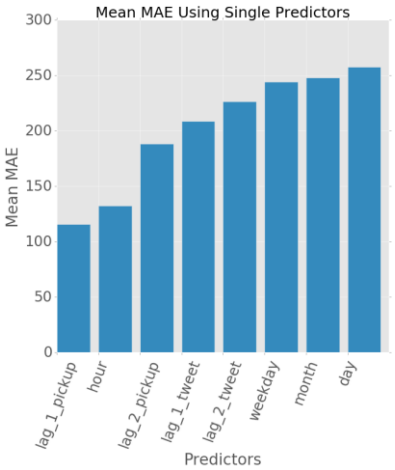

Figure 4. Using just a single feature for prediction is one way to assess which features work best.

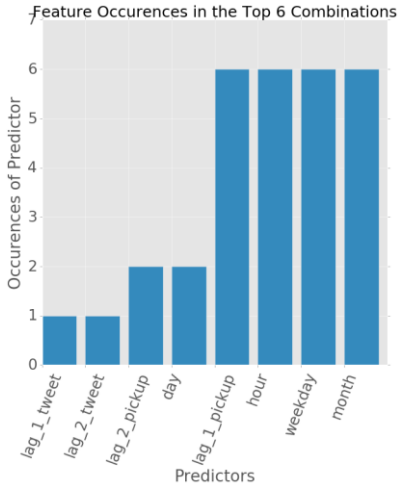

Figure 5. For the top six combinations of features, this is how often each feature was used.

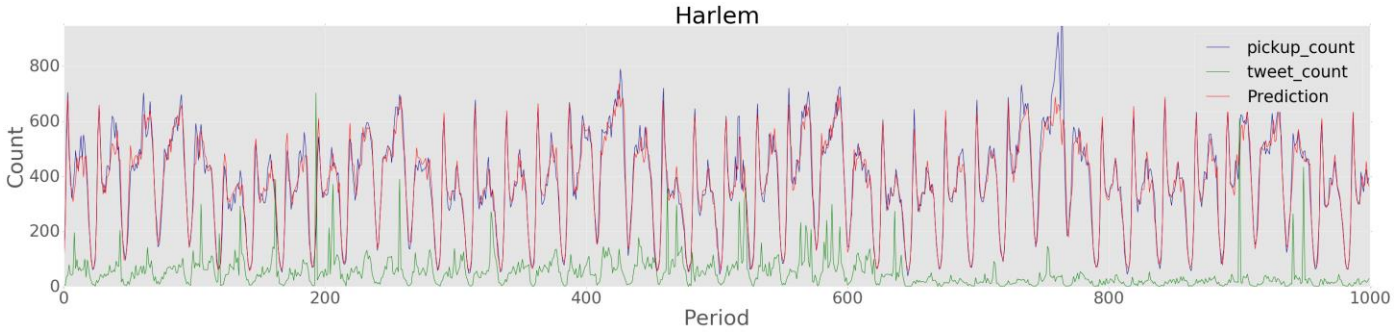

Figure 3. These time series show the hourly pickups, tweets, and predicted pickups in the Harlem neighborhood.

Additionally, we plotted the first 1,000 hours for a selected neighborhood. The plot in Figure 3 shows the data for Harlem. The blue line shows the pickup count over time, while the green line shows the tweet counts over time. We notice they tend to follow a similar pattern by time of day with some anomalies. This is to be expected as they can both be used as measures of human activity. Finally, the red line in the plot indicates the predicted values for pickup counts using all variables. It very closely follows the line for pickup counts with the majority of its inaccuracies around large spikes or isolated events.

After determining which predictors would work best, we experimented with using different combinations of predictive features for each neighborhood. This was after noticing that some neighborhoods could be finely tuned and would result in better results with certain combinations of variables. Using only the first two lag values for pickups and tweets, this final experiment was an exhaustive test of all 255 possible feature combinations for the 29 neighborhoods.

Given this data we created the two plots, Figures 4 and 5. Figure 4 shows the mean MAE for all the neighborhoods using only single predictors. This is a good way of summarizing the impact each of these has on the model. This shows that the best single predictive feature was the number of pickups in the previous hour. The next best single feature was the hour of day, followed by the number of pickups two hours prior to the predicted hour. The third and fourth best single predictors were the number of tweets from one and two hours ago, respectively.

Figure 5 shows the popularity of different features when multiple features are used to make predictions. We found that taking the six best feature combinations led to each feature being used at least once. Figure 5 shows how often each feature appeared in the top six combinations. This reinforces what we learned earlier, that the pickups in the hour before along with time-related features are most useful while the others have lower importance.

From Figures 4 and 5 overall, we see that the best features appear to be the number of pickups in the previous hour and the hour of the day. This makes sense, because the hour of the day captures the periodic nature of human mobility, and the number of pickups in the previous hour helps capture deviations from this periodicity. While the number of tweets was not generally a good feature, it does have some predictive power that could help in the absence of pickup data.

Discussion

In this work, we were able to show that the number of pickups can be predicted at a specific time of day in New York City. This fact can be used in the future for forecasting demand of taxis and autonomous vehicles. We were expecting the data mined from Twitter to have a greater impact on the model accuracy however it was only a very marginal impact. A topic for future research may be to use other external factors such as weather and scheduled events.

While these models showed good accuracy, as well as the relative importance of the features, it would be important to test over the course of a calendar year rather than just the summer months. Also, each neighborhood could have its own optimal features and model. Some neighborhoods may even benefit from further subdivision into smaller regions to improve predictive accuracy.

Conclusion

The main question for this research was to ask if taxi pickups can be predicted based on time and other data. We found that this demand can be accurately predicted, which may prove to be useful in the future when on-demand transportation is managed algorithmically in autonomous vehicles. We also showed which features are relatively better at making these predictions.

Acknowledgments

Omitted for review.

References

1. Andrew L. Kun, Susanne Boll, and Albrecht Schmidt. 2016. Shifting Gears: User Interfaces in the Age of Autonomous Driving. *IEEE Pervasive Computing,* vol. 15, pp. 32-38,
2. *NYC Taxi and Limousine Commission Trip Record Data*. Retrieved June 9, 2017 from http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml
3. Nivan Ferreira, Jorge Poco, Huy T. Vo, Juliana Freire, and Claudio T. Silva. 2013. Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips. *IEEE Transactions on Visualization and Computer Graphics,* vol. 19, pp. 2149-2158.
4. Umang Patel and Anil Chandan. 2010. NYC Taxi Trip and Fare Data Analytics using BigData. Retrieved June 9, 2017 from https://www.researchgate.net/publication/287205718_NYC_Taxi_Trip_and_Fare_Data_Analytics_using_BigData
5. John C. Krumm, Andrew L. Kun, and Petra Varsanyi. 2017. TweetCount: Urban Insights by Counting Tweets. Submitted to PURBA 2017.
6. *Pediacities NYC Neighborhoods*. Retrieved June 9, 2017 from http://catalog.opendata.city/dataset/pediacities-nyc-neighborhoods