Analysis of Chicago Trips dataset

CSP 571 Final Report

Jean-Charles Louis	Raphaël Cohen	Pranav Behari Lal
jlouis@hawk.iit.edu	rcohen6@hawk.iit.edu	plal2@hawk.iit.edu
A20431556	A20432518	A20417025

Abstract

App based taxi services such as Uber have increased in popularity over the last 5 years. This has in turn impacted more traditional taxi hailing services in many large cities around the world. We wanted to study this decline in Taxi services further, and also provide remedial options that could optimize the cost of operations for taxi companies. We have used the Chicago Taxi Trips dataset, which consist of all 120 millions taxi trips in Chicago between 2013 and 2017. To handle such a large dataset, efficient feature engineering was combined with the use of PostreSQL, to allow quick queries on the entire dataset. Initial analysis was performed to observe the trends in taxi usage, and outline possible causes for this decline. After this, two remedial options were studies: TaxiPool, which aims to quantify if ride sharing is useful for taxis, and ProfitMaximiser, which aims to predict regions of high taxi demand.

1. Overview

With the rise of intelligent transportation systems, e-hailing services (app-based services), have recently become popular among transportation users. Companies, such as Uber, Lyft, Juno, Gett, or Via, that passengers can request rides from phone application are called e-hailing service companies. The market for these services is growing, attracting more customers, and competing fiercely with other ride-hailing services in big cities in the United State, like New York City (NYC). In New York City, the number of trips by street hailing taxis (yellow cabs) has fallen between 2014 and 2015, while, during the same time period, the demand for e-hailing companies such as Uber has increased significantly.

Drivers in hailing services (either e-hailing or street hailing) have to search for their next passenger, which entails driving an empty taxi around the city. At the same time, in some parts of urban areas, passengers may have to wait for a long time to find a cab. It has been shown that having a better knowledge of demand in the near future can improve the efficiency of the system [1]. It can help drivers reduce their empty cruising by suggesting locations where they might find passengers [2] and help passengers reduce their waiting time. Moreover, decreasing empty taxi travel time and distance can result in less congestion and pollution [3].

In this research, two methods were studied to improve the efficiency of taxi usage in Chicago. First, 4 years of taxi trips from 2013-2016 were studied to understand the feasibility

of ride sharing in Chicago. The idea is, ride sharing would drastically reduce the price per mile for customers, and would allow taxi companies to better manage their taxis, given the reduce in demand over the years. Traditional clustering methods like K-means and DBSCAN clustering were considered along with a manual grouping algorithm based on SQL. Pros and cons of each method were identified.

Predicting the demand in different regions of the city, at different times of the day, can help balance the supply of taxis around the city. This can both reduce the empty riding time for taxi drivers, and also reduce the average wait times for customer. To do so, literature pointed towards the ARIMA model which was used in the analysis. Variations of the ARIMA model were compared to arrive at a final model to predict demand in the 25 regions in the Downtown Chicago regions, with predictions being made in 15 minute intervals through the day.

The Chicago taxi trips data from the Chicago taxi portal was used for the analysis. The dataset consists of roughly 120 million taxi trips in Chicago from 2013 - 2017. The dataset consists of features such a pickup and dropoff locations, price, duration, and distance covered by each trip.

---

2. Data Processing

Taxi trips reported to the City of Chicago in its role as a regulatory agency. To protect privacy but allow for aggregate analyses, the Taxi ID is consistent for any given taxi medallion number but does not show the number, Census Tracts are suppressed in some cases, and times are rounded to the nearest 15 minutes. Due to the data reporting process, not all trips are reported but the City believes that most are.

2.1. Initial Preparation

Our dataset is 43Gb with 113 million trips and 23 features. Some feature are not encoded in an efficient way:

- The start time is split across multiple feature (day, day name, month…). We group them into a single timestamp, easier to work with.
- Most numerical features where encoded as strings (eg Paid amount was a string with dollar sign at the end). We converted them to pure numerical.
- Removed unnecessary feature like trip_id which is an internal 256bits representation that takes space without providing something useful.
- Removed redundant features.

Making these changes has reduced the size of our dataset to ~18 GB.

We also split the data into multiple smaller files, which made the initial analysis easier:

- Year files
- Weekly files
- Random sampling across the dataset

2.2. PostgreSQL

This size reduction and re-encoding was not enough to make working with the whole dataset easy. We ended up using PostgreSQL, as suggested by the professor.

This prove really useful and easy to work with in R. We were able connect to the local database, make SQL query in R and load the result in a dataframe. PostgreSQL allowed us to make queries on the whole dataset, which was impossible with R datatables.

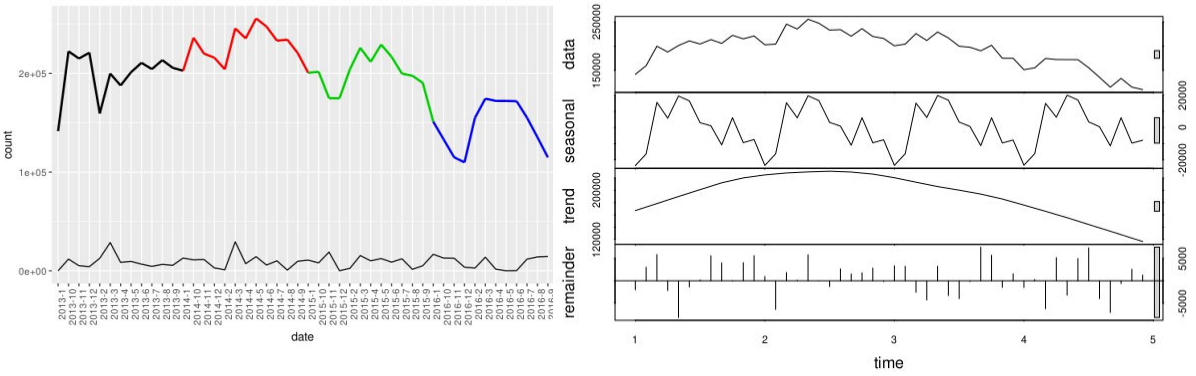
---

3. Data Analysis

"Chicago cabbies say industry is teetering toward collapse" - [4]

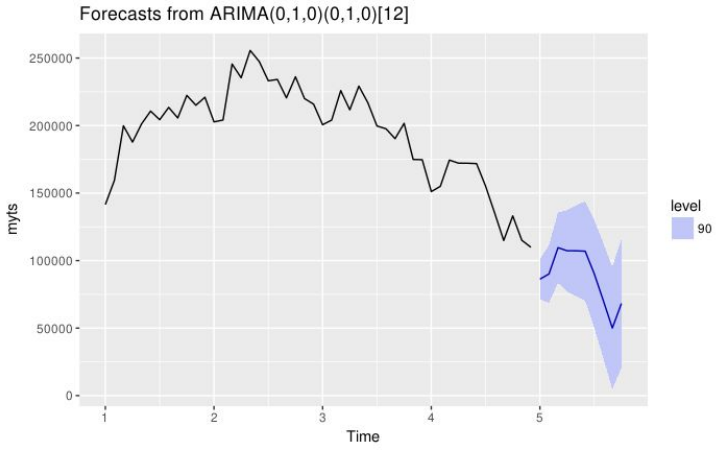
What a pessimistic claim ! Our main focus for this part has been to investigate whether these reports were true or not.

The most natural way of looking at this problem is by considering the evolution of taxi usage throughout the years. Below are shown two graphs, respectively, taxi usage by month from 2013 to 2016 (one color per year) and its time series decomposition. It is worth noting that we do not have data past september 2017, which means that these two plots above stop before 2017.



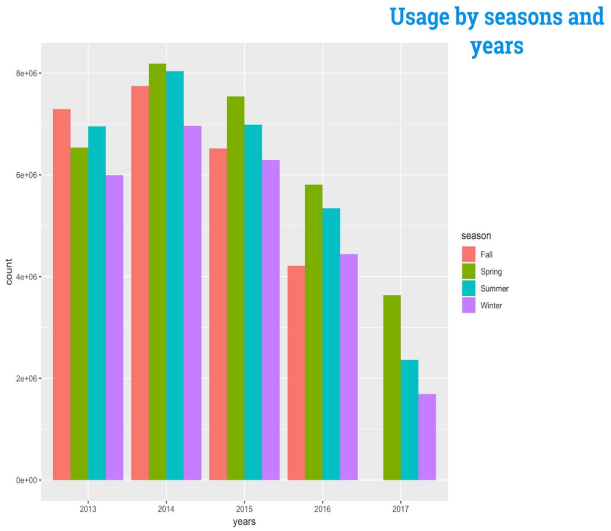
It is pretty clear that there is a reduction in taxi usage over the last 4 years.

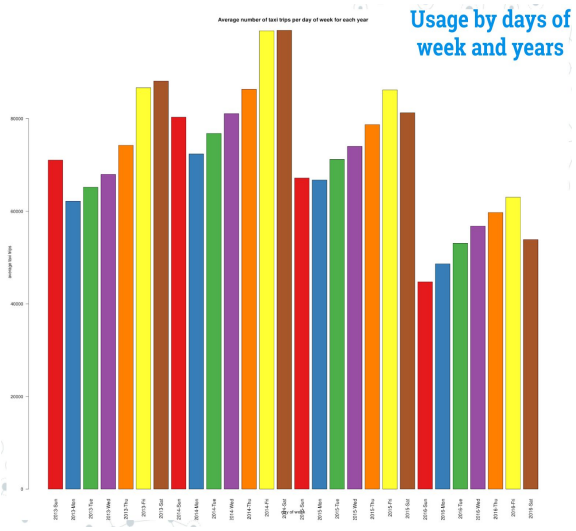
We tried to predict what would happen for the next 10 months with an ARIMA model. The figure on the right appear to be validating what we thought. Even though the prediction might be extreme and very pessimistic, the trend is still downward.



For the following, we will take a look at taxi usage evolution over the seasons as well as the day of week.

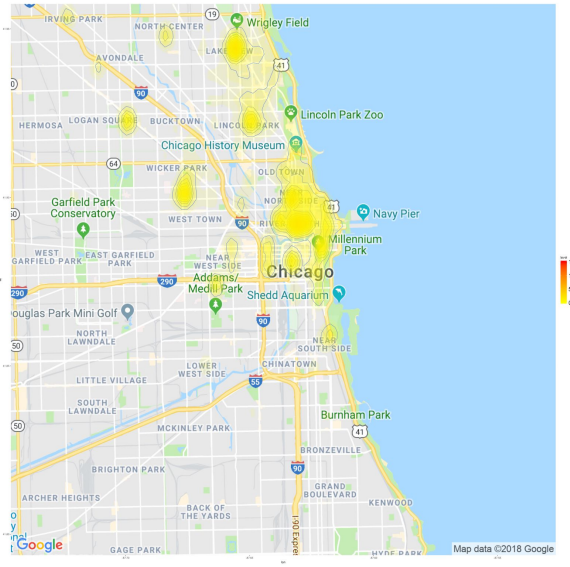
These plots, once again, confirm our previous results.

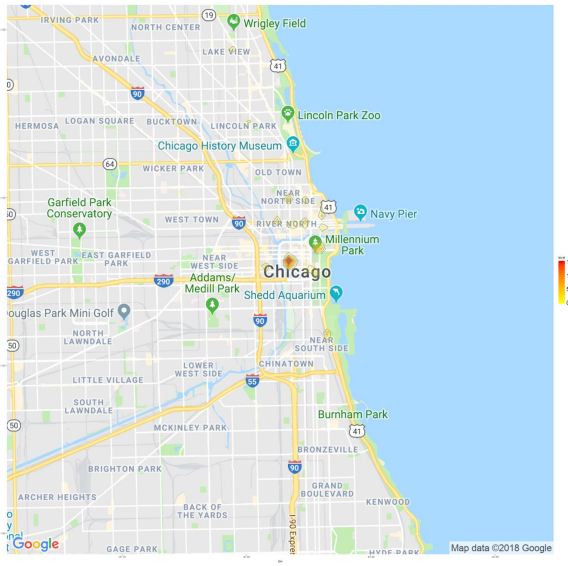




The barplot on the right show that taxi usage is steadily increasing from monday to saturday where it reaches its maximum, and decreases from saturday to monday. However, we can see that over the years the average numbers per day are declining.

Now that we have seen trends and predictions for taxi usage from 2013 to 2017, it is time to scale down and observe taxi density by hours on a chicago map. As this is a static document, it is impossible to include an animated GIF, but we will try to extract general tendencies from those two maps only.

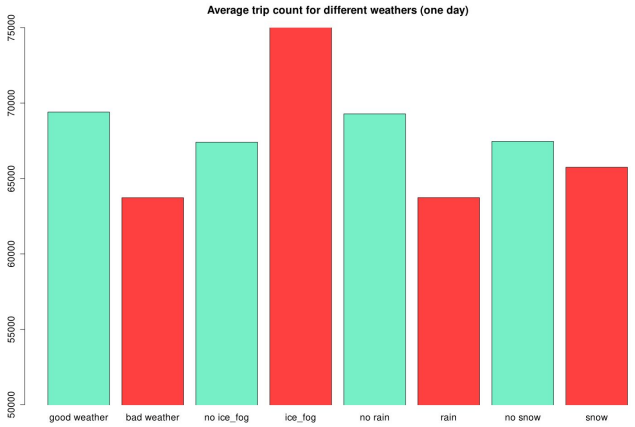




The map on the left represents extremely well the taxi density level during the night as well as during the afternoon for pickups and dropoffs. The traffic is highly spread out over the city but not dense.

On the other side, we have an opposite representation that mimics taxi usage for the morning and evening. It is concentrated around one major point that is the extreme center of the Loop. It is not showcased on those two plots above but it is interesting to mention that for the pickups, the map is more dense during the evening than the morning. It is the other way around for drop offs.

Here is a barplot of taxi usage by weather. We can see that, on average, when the weather is good, people tends to use the taxi more. With the exception of “ice fog” which doesn’t seem to scare people enough to stay home, but still might hold them back from walking.



4. TaxiPool

Data from Google Trends showed us that searches for “Uber” in Chicago are rising. On the same time period, the number of taxi trips in Chicago is decreasing. Even is correlation does not imply causality, Uber could be one of the causes of the decrease of usage in taxi.

One clear advantage that Uber has over taxis is the “Pool” feature that group people doing similar trips in the same car. This allows for cheaper trips. The goal of TaxiPool is to decide if ridesharing is possible for taxis in Chicago.

4.1. Model solution

Our first idea was to use an existing clustering solution. We wanted to find cluster of similar trips. Finding a lot of those clusters would prove the usefulness TaxiPool.

Clustering would be done in 5 dimensions: 2 for both pickup and dropoff (latitude and longitude) and 1 for time. We looked into existing solutions: K-Means and DBScan. Understanding their inner working and trying them on sample data showed us that those algorithms are meant to discover patterns. Furthermore, they don't fit a case where you want to limit the size of the cluster: DBScan has a min points, but the cluster can grow infinitely large as long as they are trips less than one epsilon away. This is not acceptable for our problem because we want to clusters trips with pickup and drop-off within a walking distance. Also, we have no interest in discovering patterns, we are on the contrary looking to see if the data fit a particular predefined pattern.

4.2. Algorithmic solution

So we realized that a manual algorithmic solution is a better way to solve this problem. Indeed, we wrote everything in one big SQL query.

The first step was to filter trips that we don't want to include in this analysis: trips without GPS data and trips that where we cannot trust GPS data (see next paragraph). We then round latitude and longitude to 3 decimal places, in order to put trips in a grid of about 100m by 100m squares (in fact 88m by 111m because one unit of latitude is not the same as one

longitude). We then group trips with the same start time, rounded pickup and drop-off. The final step is to count how many group they are for each group size.

We also modified it to output the pickup, drop-off and start time of the groups with the most trips.

4.3. Results

This algorithm runs successfully in under 12 min on a laptop over the whole dataset.

It first made us realise there are some odds trips in the dataset. Worst exemple is 19 Aug 2014, where in 30 min interval there were:

- 183 trips
- all going to and from the exact same point
- ranging from 0.2 to 17 miles

It is clear that we can not trust the GPS coordinates from those trips. We decided to exclude all the trips where the pickup and drop-off locations are the same. We discuss a better way to remove trips where GPS data doesn't make sense in the future work section of this paper.

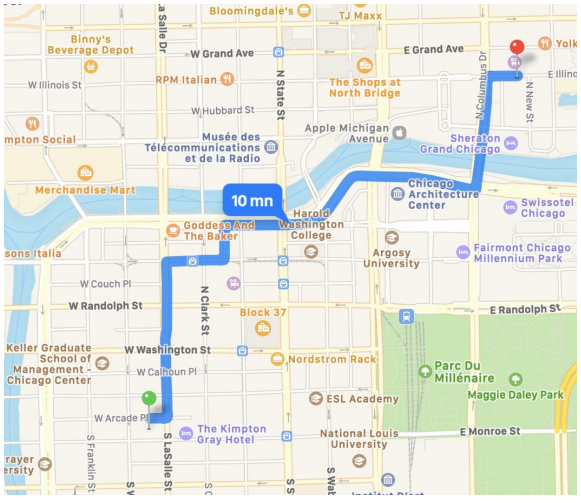
The results were very good. 53% of trips can be group with a similar trip within 15 min. Out of those 53%, 17% can even be group with more than 5 other trips.

However when need to keep in mind that those results may not reflect the full reality. Indeed, while we know that times are reported in intervals of 15 min, we have no precise information about the GPS data. Even if they are provided with a really high accuracy in terms of decimals points, it is very likely that they were rounded in one way or another that made trips seems closer than they really are.

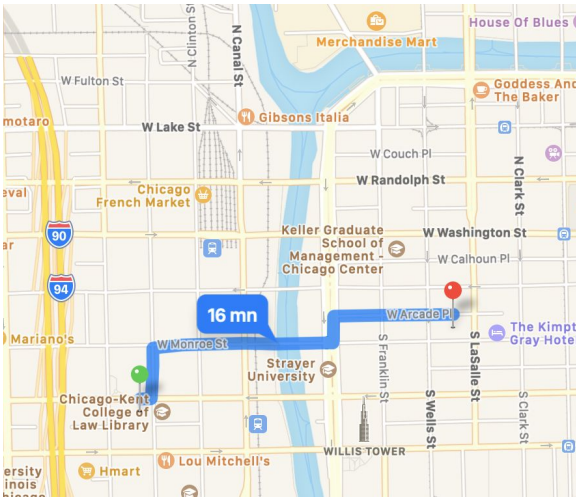
Extremes

We also found some interesting edge cases in the results. Below is the two cases where the polling was most successful. In both cases, we have 61 trips in a 15 minutes interval that are really close to each other.

Feb 11 2014 12:30



Feb 19 2015 08:30



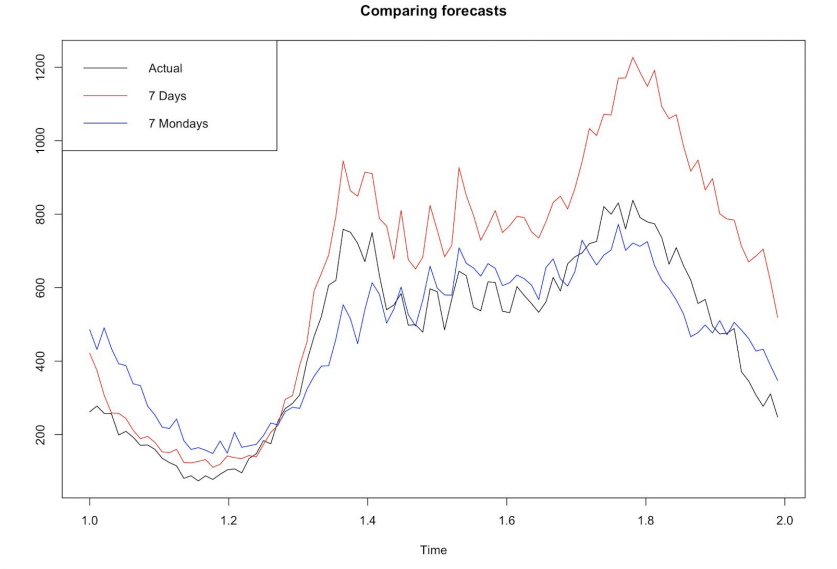
5. ProfitMaximiser

A major reason for the decline of taxis could be attributed to the rise of Uber. A simple correlation between the increasing popularity in Uber and the decline of taxi usage supports this claim. The advantage with Uber lies in its app-based approach. By knowing when and where their app is being used, Uber is able to better manage the supply of cabs to high demand areas. For the case of Chicago taxis, such demand prediction can only be made on the basis of historical data. [5]

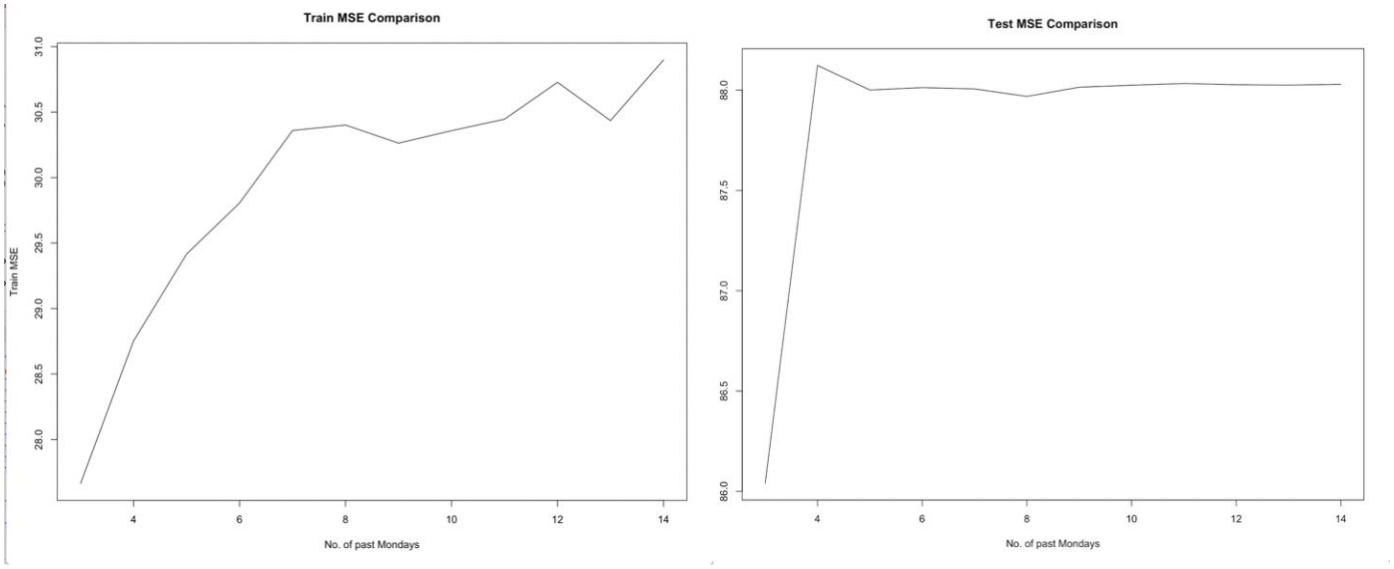
Owing to the potential cost saving, plenty of research has been done to predict the demand of taxis in several cities. Initial research about ride-hailing for NYC taxis was undertaken in a study done by Schaller [6]. To understand the behavior of taxi demand, the study applied a citywide empirical time series regression model in order to better picture the relationship between taxicab revenue per mile and economic activity in the city, taxi supply, taxi fare, and bus fare. Moreira-Matias [7] developed an approach to predict short-term taxi demand at 30 minute intervals. Their approach compared three predictive models, a Time Varying Poisson Model, a Weighted Time Varying Poisson Model, and an Auto-Regressive Integrated Moving Average Model. It was found that the ARIMA model performs better than the former two. To train the models, two days of historical data were taken and used to predict the short term demand in 30 minute intervals. This approach was slightly modified in our work, with the final predictions made using an ARIMA model.

The Chicago taxi dataset rounds of each taxi trip to the nearest 15 minute mark, and so, our models aim to make short term demand predictions in 15 minute intervals. The paper mentioned above assumed the last 2 days of historical data to give a good indication of taxi usage demand. To put this to the test, two alternatives were tested. First, it was checked if the past *n* days of historical data would give an appropriate fit, or if taking the past *n* week days be more appropriate. That is, if we are trying to predict the demand for a Monday, would considering the last *n* days give better predictions, or would considering the previous *n* Mondays. To put this to the test, two ARIMA based models were fit with training data involving either the previous 7 days, and the previous 7 Mondays.

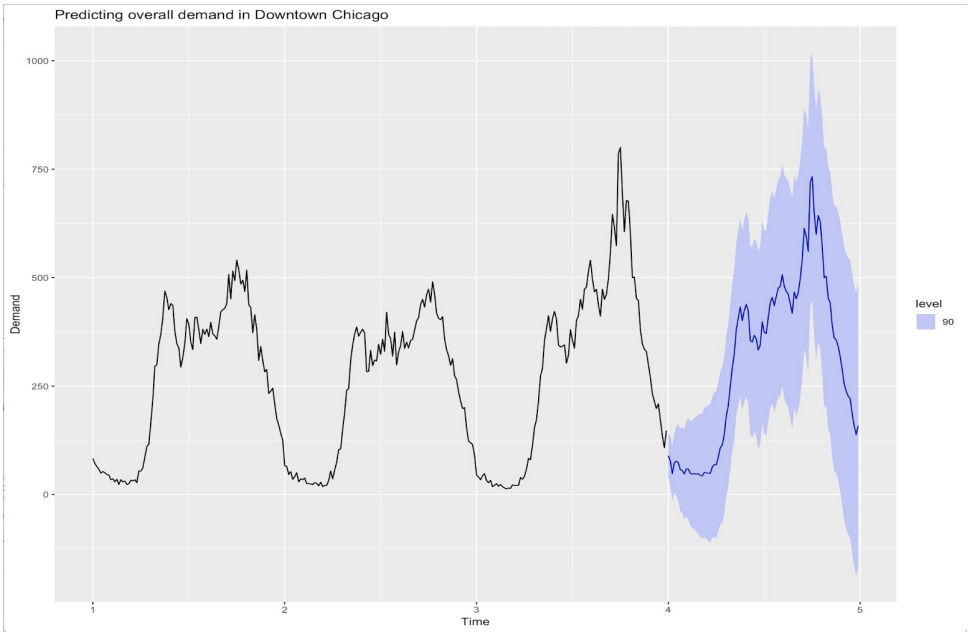
Comparing the demands shows that the previous 7 Mondays gives a much closer fit to the actual data. The horizontal time axis denoted 96 15-minute intervals through the day.



Once a decision was made to use the previous Mondays to predict the demand for a given Monday, the next question to answer was how far back do we go. Models were run to test taking 3 to 14 historical days, and compared for their Mean Squared Errors (MSE) on the training and a test dataset.



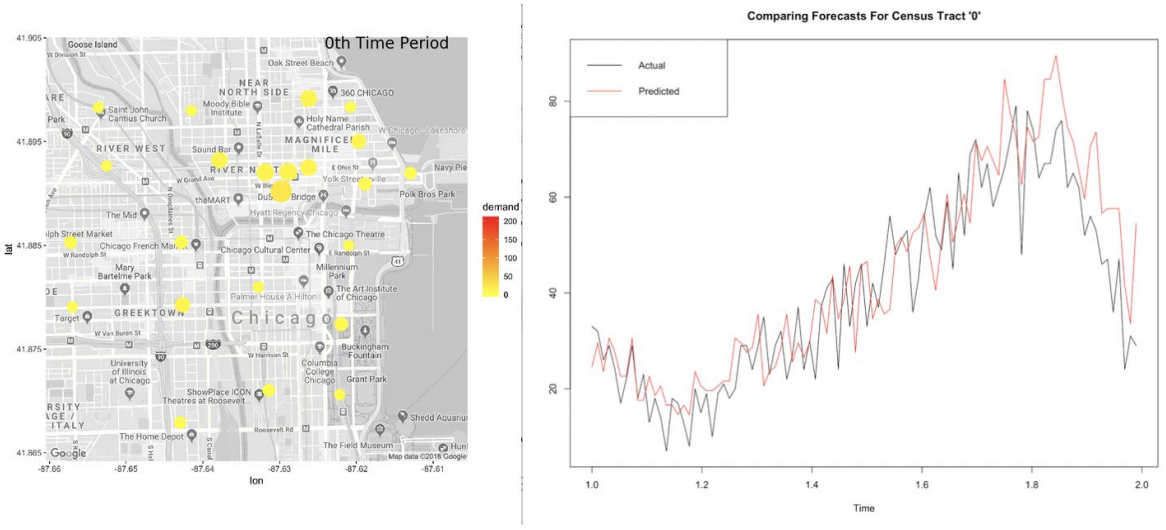
The results showed a stark drop in performance with 4 more or more days of historical data. The model considering the last 3 Mondays was found to be optimum.



To reinforce our ideology of considering previous Mondays instead of considering the 3 previous days, the two models were tested once more, and the model considering previous Mondays performed far superior.

So far, the predictions made have been the city of Chicago as a whole. To have any value to either taxi companies or its drivers, it would be beneficial to know which specific area to target. To simplify the implementation of our idea, we have only focused on the 25 regions defined by Census Tract in the Chicago downtown area. Our dataset was filtered down to include trips in these regions only, and any trips with missing pickup or dropoff locations

were omitted. The ARIMA model considering the last 3 Mondays was fit once more, but this with data for each Census Tract individually. That is, for the 25 regions under consideration, we have fit 25 different ARIMA models. For example, the training set to predict the demand on a Monday at a region with Census Tract A included all trips that had a pickup from the region in Census Tract A for the previous 3 Mondays. The predictions made for each individual Census Tract were compared with the test dataset, and were found to give a good fit. The models created are able to predict the demand in the 25 regions of the Chicago downtown area in 15 minute intervals.



6. Conclusion and Future Work

As we realised in the TaxiPool part of the project, some trips GPS data don't really make sense. We found a significant number of them going to and from the exact same point but with way different reported length in miles. We excluded them, but a better solution would have been to compute the straight line distance (or even better with Manhattan distance) between those two points and compare it to the reported trip length in order to exclude trips where it's too far for the datapoint to be trusted.

For what we call 'Profit Maximiser', we analyzed several variants of the ARIMA model to make meaningful predictions of the taxi demand. First, it was found that for a particular day of the week, considering taxi trips from that day (like Monday) from the previous 3 weeks gives a better indication of the expected demand that does simply taking the previous 3 days. This altered model was then used to predict the demand in the 25 regions in the Downtown Chicago area in 15 minute intervals.

For future scope on the project, we would like to explore the correlation between taxi usage and weather further. Also, it would be worth correlating taxi trips and the routes followed with the other forms of Chicago's public transport like the 'L' or buses. These could give us a better indication of the usage of public transport in Chicago.

Data Sources

The main dataset is “Taxi Trips” from City of Chicago Data Portal.
https://data.cityofchicago.org/Transportation/Taxi-Trips/wrvz-psew/

Source Code

We used GitHub for collaboration.

Github Link: https://github.com/louisjc/CSP571-Project.git

Citations

[1] Sabiheh Sadat Faghih, Abolfazl Safikhani, Bahman Moghimi, Camille Kamga, ‘Predicting Short-Term Uber Demand’, arxiv.org:1712.02001

[2] Wang, D., Cao, W., Li, J., & Ye, J. (2017). DeepSD: Supply-Demand Prediction for Online Car- Hailing Services Using Deep Neural Networks. Data Engineering (ICDE), 2017 IEEE 33rd International Conference on (pp. 243-254)

[3] Li, Y., Lu, J., Zhang, L., & Zhao, Y. (2017). Taxi booking mobile app order demand prediction based on short-term traffic forecasting.Transportation Research Record: Journal of the Transportation Research Board, (2634), 57-68

[4] Aamer Madhani, https://www.usatoday.com/story/news/2017/06/05/chicago-cabbies-say-industry-teetering-toward-collapse/102524634/

[5] Qian X., Ukkusuri S.V., and Yang C., Yan F. (2017). A Model for Short-Term Taxi Demand Forecasting Accounting for Spatio-Temporal Correlations. Transportation Research Board Annual 2017, Washington D.C.

[6] Moreira-Matias L., Gama J., Ferreira M., Mendes-Moreira J., and Damas, L. (2013). Predicting Taxi-Passenger Demand using Streaming Data. IEEE Transactions on Intelligent Transportation Systems, Volume 14, Issue: 3, DOI: 10.1109/TITS.2013.2262376

[7] Yiming, Wu. 2016 Chicago Cabs Analysis, 2016.

[8] Todd, Schneider. Chicago’s Public Taxi Data, 2017

[9] Uber. Chicago: A Uber Case Study, 2015

[10] Regulatory Reform Team - Ash Center at Harvard Kennedy School. Case Study: New York City Taxis, 2014

[11] Y Lin, W Li, F Qiu, H Xu. Research on Optimization of Vehicle Routing Problem for Ride-sharing Taxi, 2012