Optimizing the Location and Use of Taxi/FHV Relief Stands

Huy T. Vo[1], Kaan Ozbay[2], CUSP capstone manager Federica Bianco[1], Cheng Hou[3], Vishwajeet Shelar[3], yw2278[2], and Le Xu[3]

[1]NYU Center for Urban Science & Progress
[2]Affiliation not available
[3]New York University (NYU)

August 3, 2017

Abstract

A decision support system for placing taxi/fhv relief stands must be implemented because of the surge in for-hire-vehicles (fhv), though the decision baseline of where to put a taxi relief stands is rather blurred. This study is focused on exploiting an understanding of existing taxi relief stands usages, taxi ridership, parking violations, and both geographic and demographic factors, to tackle this problem and set the certain ground for future decision making. Various data sets are collected throughout the analysis, including the public park restroom data, hotel data, restaurant data and parking violation (tickets) data. For this single project, over 300 GBs of data is processed by streaming in Spark. Individual taxi GPS locations and trips over the Year 2015 are parsed and the relatively highly-used and less-demanded relief stands are identified. A Rank Function and a Mixed-Integer Linear Programming model are implemented with a focus on minimizing the parking violations in terms of numbers of projected taxi relief stands. Another model based on the demand of customers and need of a taxi relief stand is also implemented. A ranked taxi relief stands recommendation among the 4003 hexagons, as well as a comprehensive statistical report on the relief stand usage is provided. The importance of this work is far beyond allocating taxi relief stands, within the urban context, a similar approach could solve problems such as allocating urban facilities, for instance, electric vehicle charging stations, gas stations, distribution centers, or subway stations.

1 I. Introduction

2

2.1 Background & Motivation

"There are over 100,000 licensed For-Hire Vehicles (FHVs) in NYC. The For-Hire industry is constantly growing with more vehicles going out on the road every day. An increase of vehicles on the road has led to more congestion on the road. Some drivers may take short breaks from work and inadvertently add to this congestion. In an effort better manage the curb space, taxi/FHV relief stands have been implemented to give drivers a place to pull over and take a break. However, there are only 69 taxi Relief Stands (TRSs) spread across NYC, in addition to the taxi stands allocated for the active pickup and drop-off of passengers. " (USI 2017 Capstone Project Catalog)

The aim of the project is to look at the effectiveness of current Taxi/FHV Relief Stands as well as identifying possible locations for new Taxi/FHV Relief Stands (TRSs) and also develop a scientific decision model that helps present valuable insights to the city planners and traffic engineers in answering questions such as how

to position relief stands; where they are most needed and where they would not burden the traffic at the same time. This project's concept is proposed by NYC Taxi and Limousine Commission in order to help better the lives of taxi drivers, and various taxi related data sets are provided to fuel this study. The NYC taxi data is one of the largest data sets available to the public, big data technologies such as Hadoop and Spark are used to fulfill the expensive computational requirement.

3 II.Related Work

There has been a lot of research on allocating optimal "service providing" locations for taxis. Most of the research is concentrated to maximize the use of taxis by making taxis available to the customers as soon as possible. Our research is about finding locations for taxi relief stands and is more concentrated towards maximizing the usefulness of the stands for the taxi drives instead of the customers. Even though the target audience for the taxi stand may be different, the applied methodologies are quite similar.

The problem of allocating locations for taxi relief stands has been solved using both qualitative as quantitative methods. The Taxi Rank Master Plan for Melbourne City which was prepared by MRCagney Pty Ltd (MRCagney Pty Ltd, 2012) provides a qualitative guideline for factors to be considered while placing a taxi stand. The report discusses the importance of factors like proximity to physical locations like intersections, downtown areas. It also provides guidelines about the technical factors like the width of a road, the curbside restrictions. Since the report is developed for providing optimal taxi stand location for use of the customer, it does not incorporate the importance of factors like proximity to restrooms, restaurants which are to be considered for taxi relief stands.

The facility location allocation problems have been used in other sectors as well. One of them is the electric car charging stations allocation problem. The paper "Optimizing charging station locations for urban taxi providers" proposes "a decision support system for placing charging stations to satisfy the charging demand of electric taxi vehicles." **(J.2016)**(J. Asamer et al., 2016) The study is based on the implementation of charging stations in the city of Vienna, Austria. The model proposed by the authors provides a set of areas (hexagons) where charging stations can be implemented. The charging points are chosen to minimize the distance between the pickup location of customer and charging station. Other factors are like existing charging stations, the estimated demand are also used to solve a linear optimization problem. Our research uses a similar approach of creating hexagonal tessellations. Calculating a cost function for each hexagon using factors that are to be considered for a relief stand.

4 III. Data Source

The project is built on two major datasets. The first one is the Breadcrumbs data, which has records of the GPS coordinates of all the yellow and green taxis at every two minutes. The Breadcrumbs data is categorized as Yellow data (not open to the public), and it is authorized to CUSP students for certain taxi data related projects. For further disclosure security, the personal information from the breadcrumb data is transformed by Professor Huy, Vo. The second dataset is the Taxi Trip data, which contains millions of trip records for both Yellow and Green taxis. The trip data include both pickup and drop-off points, which can be used to understand taxi demand within a certain geospatial unit. The combined size of these two datasets is about 300+ GBs.

Other datasets like restrooms in public parks, restaurants, hotels, parking violations and relief stand request data are also used.

CUSP data hub
- Taxi Trip Data (Yellow and Green Taxi) - 2015
- Taxi Breadcrumb Data (Yellow and Green Taxi) - 2015 - Not to Public

NYC Department of Transportation
- Taxi Relief Stand Locations

NYC Open Data
- NYC Park Restroom
- Parking Violation Issued - 2015
- Parking Violation Codes

NYC Restaurants and Hotels
- Venues Data

TLC Internal Driver complaint information (TLC Internal Data)

5 IV. Methodology

6

7

The goal of the project is to quantify the effectiveness of current Taxi Relief Stands and identify possible locations for new Taxi/FHVs Relief Stands (TRSs). Firstly, the breadcrumb and the public taxi trip records are examined to identify the TRSs as well as the Non-TRSs locations where most drivers spend a significant amount of time. Besides the breadcrumbs and taxi trips datasets, the restaurant, public parks, hotel, and the parking violations datasets are also utilized to understand the popular places among the drivers. The expectation of the TLC is not to find an exact location for new Taxi/FHVs relief stand, so a block unit (hexagon) is suggested.

7.1 Data Acquisition/Preparation

- **Hexagon Grid**

The client wanted the suggested locations to have a size of about 4-5 blocks. A constant geographic zone was created by dividing New York City into Hexagon Tessellations, each hexagon was called zone or region. Each zone roughly corresponds to 1.6 km perimeter. In total, NYC was divided into 4003 hexagonal zones.

Regular hexagons were chosen to create the regular tessellation because they are the closest shape to a circle. They also provide reduced edge effects and all the neighbors are identical. (Matt S. Mackey, 2016)

- **Taxi Relief Stands**

There are 69 TRSs listed on Department of Transportation's official Website. The data provided by the DOT lists 62 relief stands. Further discrepancies were observed after checking the relief stands by physical visits and virtually checking on google street view. The discrepancy of the data was brought to the notice of the client (TLC) and they agreed upon using the DOT dataset which is available. The deviation of the data is about 5-6 relief stands, which is within the acceptable limits of the TLC. The figure below shows the map of the location of current taxi relief stands.

Figure 1: The pink spots represent the current taxi relief stands from the data given by DoT. There is total of 62 relief stands available out of 69, and 45 among them are in Manhattan.

- **Park Restrooms**

Parks are great facilities designed to enhance the life qualities of urban dwellers, and to place TRSs around parks equipped with toilets is an optimized way to maximize land use. This project examined the public parks which provide restrooms. The number of park restrooms is an important factor when placing potential relief stands.

- **Parking Violation**

The year 2015 Parking Tickets/Violations record is used in this project. There are 32617 parking violations issued to TLC licensed vehicles (Taxi/FHV). These violations are related to parking only, they do not include the speeding and other types of non-parking violations. A positive relationship is observed between the parking violations and the top used relief stands. This makes parking tickets an important feature because it directly impacts the income of the drivers.

- **Restaurant and Hotel Data**

In the project scope, the importance of restaurants is crucial as restaurants provide both food and toilet for taxi drivers. However, the restaurant data could be enormous due to the characteristics of New York City, a city with more than 30,000 restaurants (MenuPages statistics). According to the TLC Fact Book (TLC, 2016), the top five origins (place of birth) of the taxi drivers are Bangladesh, the Dominican Republic, the United States, Pakistan, and India. Since more than 40% of the drivers that are originally from Bangladesh, India, and Pakistan, both Indian and Bangladeshi restaurants are part of the feature space. All different ethnic groups are included in the study by considering the less expensive restaurants that have entree price less than $10. This helps to minimize the potential unfairness impact that this analysis approach could bring.

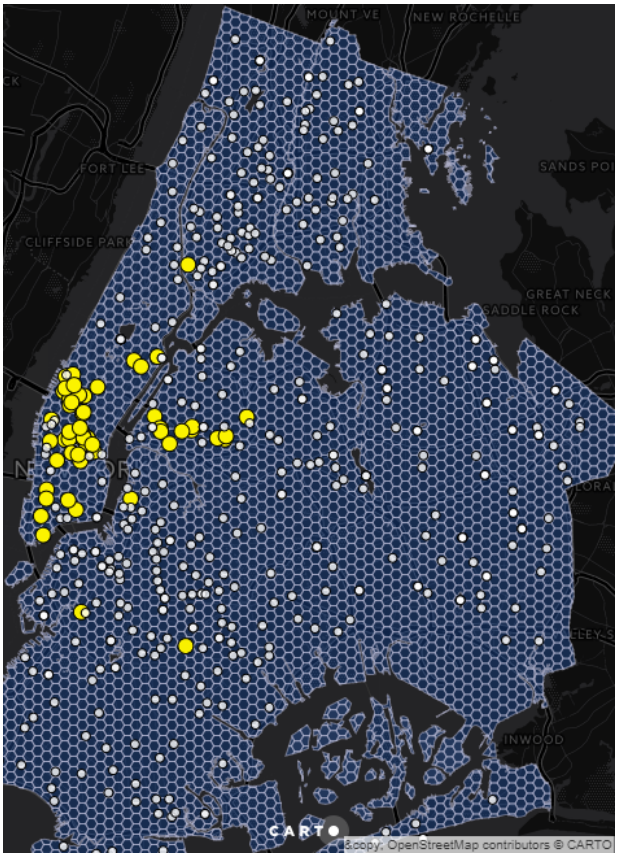

Figure 2: In this plot, the yellow dots presented the current taxi relief stands, and the white dots are the public parks with restrooms. This plot offers an understanding of how the relief stands and parks are located among each other.

Hotels are being considered as one of the important features in this study. As information gathered from several drivers, they normally would use hotel loading zones as a "relief stand", the reason is that they can use the bathrooms from the hotel and also pick up potential hotel customers on the spot, disregarding the possible parking violations. The hotel becomes a feature when considering where to put new taxi relief stands.

Foursquare API offers access to its world-class venues database, specific restaurant and hotel information is downloaded by adding parameters, such as price, location, category, and radius range. API calls are made for both current relief stands analysis as well as the overall New York City zoning analysis and lists of nearby restaurants and hotels within the radius of 500 meters are obtained. Due to the call limitation and offset limit, the maximum number of restaurant or hotel counts that is obtained from each call is 50.

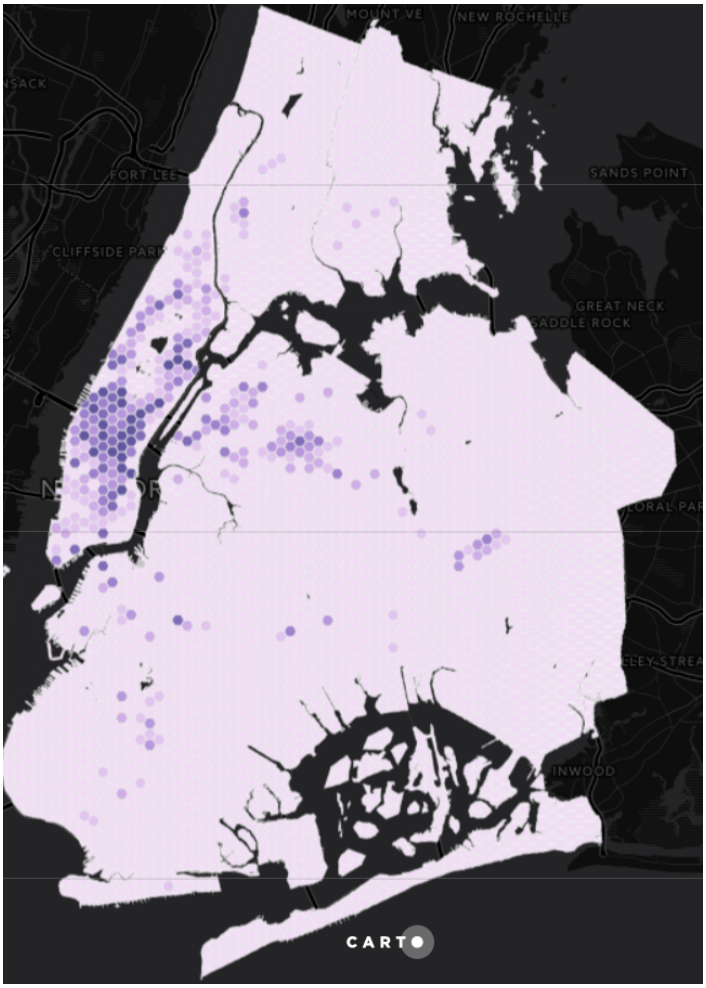

Figure 3: The plot is the choropleth map of the taxi parking violation count in the year 2015. The darker the purple hexagon represents the higher amount of parking violations counts. From this image, the violations are clustered in the midtown area.

7.2 Big Data Analysis

7.3

One of the characteristics of the project is the Big Data. TLC provided more than 300 gigabytes of taxi breadcrumb records data and trip records data. Using big data management methods and skills, all the required data is processed successfully within half an hour using the High-Performance System (CUSP cluster). The figure below helps to explain what valuable information is extracted from the big data processing.

Fig 6 shows that the breadcrumb and taxi trip data can help to answer two questions:

1. **Where is the taxi stopping for more than 30 minutes?**
2. Is the taxi stopping at the taxi relief stands?

For this particular example, only 1 out of 5 breaks is identified on an existing taxi relief stand. The detailed

Figure 4: The distribution of hotels in each hexagonal zone is displayed on this map. The hotels are mainly located in Manhattan or areas that near Manhattan. The darker the color, the denser the hotel density.

algorithm application is provided below.

- **Current Relief Stand Usage - Idle Time Aggregation**

In order to answer the two questions posed earlier, it is necessary to understand the usage of 62 relief stands and identify locations that taxis/FHVs drivers spend a significant amount of time. Firstly, the data is cleaned by filtering out the GPS points in Breadcrumbs data that occur between each trip's pickup and drop off time interval, since drivers will never use the relief stand while in business. Next, to identify all the idle points within a certain location, each vehicle from the breadcrumb data is tracked with its unique id, the speed for each taxi is calculated and the speed threshold is set to 5 (meters)/mins. All the points with speed lower than the threshold are identified as idle points and extracted. For continuous rows with total stop time over 30 mins, the centroids of these rows are taken as the final idle point. The relief stand usage is calculated by creating a buffer of 50 meters around the relief stand. All the idle points within the relief stand buffer are counted as the number of idle cars and the total use time of all the idle cars is aggregated as the idle time of the taxis. The below flow chart (Figure 3) shows our algorithm.

With the help of the distributed computing power of Spark, a large amount of computational time is saved. The idle points are obtained with Spark by processing the data in a streaming fashion. Each data point is accessed only once. A flag is created and the default value is set to 0. Next, a function is defined to check if the speed (distance/time) between two consecutive points is less than the defined speed threshold (5 m/min), and if so, the flag is set to 1. Such points are tracked for their numbers, duration, latitudes, and longitudes, till the speed gets larger than the threshold. Next check is on the gross time, is the time is larger than or equal to 30mins, the centroids of these points are outputted as idle points (idle cars) and the gross idle time is also extracted.

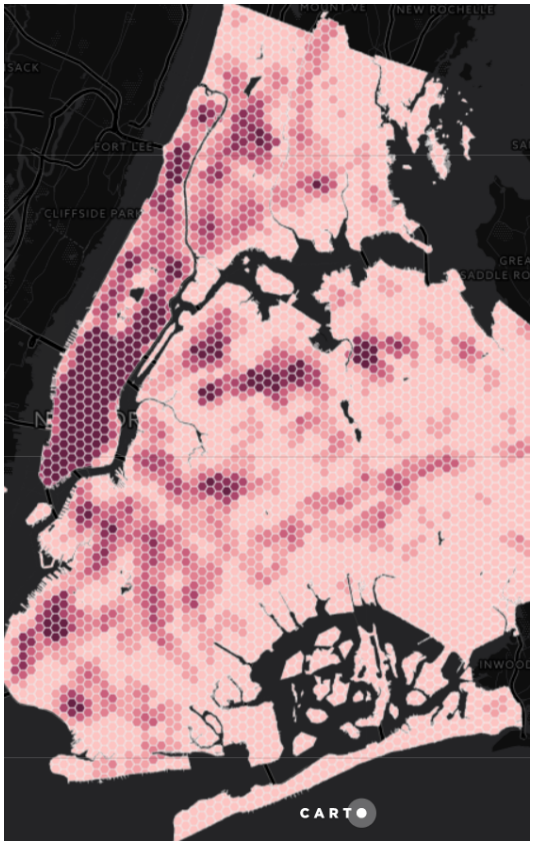

Figure 5: This map shows the numbers of affordable restaurants within each hexagonal zones. The darker the color, the denser the restaurant density

- **R-tree**

The efficiency of the spark script which maps idle points to relief stand buffers and hexagonal zones is improved using R-tree. As a tree data structure algorithm designed for spatial search, R-tree divides a geolocation into bounding rectangles, and both the bounding rectangle and its contained objects (rectangles) do not intersect each other. Within the bounding rectangle, smaller sections are further divided. By using R-tree, each idle point with coordinates could be easily and efficiently assigned to the section that it belongs to and ignore the search in other irrelevant sections. The Spark script uses R-tree algorithm to check all the point and then the distributed computing property is utilized by using the map-reduce method.

- **Taxi Trip Counts (Pick-ups, Drop-offs)**

Similarly to the idle points (taxis) count, the taxi ridership is also considered when deciding where to put next relief stands as it would be convenient for taxi drivers to access potential customers right after the break. The pick-up and drop-off times and locations are obtained from the trip records data by using MapReduce. These points are aggregated to the hexagons using the R-tree algorithm mentioned above. R-tree is used because the size of the dataset is considerably large; ˜1200,0000 points to be mapped with 4003 hexagons. Implementing normal intersection using geo-pandas is computationally expensive and slow because it uses pair-wise searching and it could take more than 10 hours to process, but with R-tree the process time is reduced to half hour.

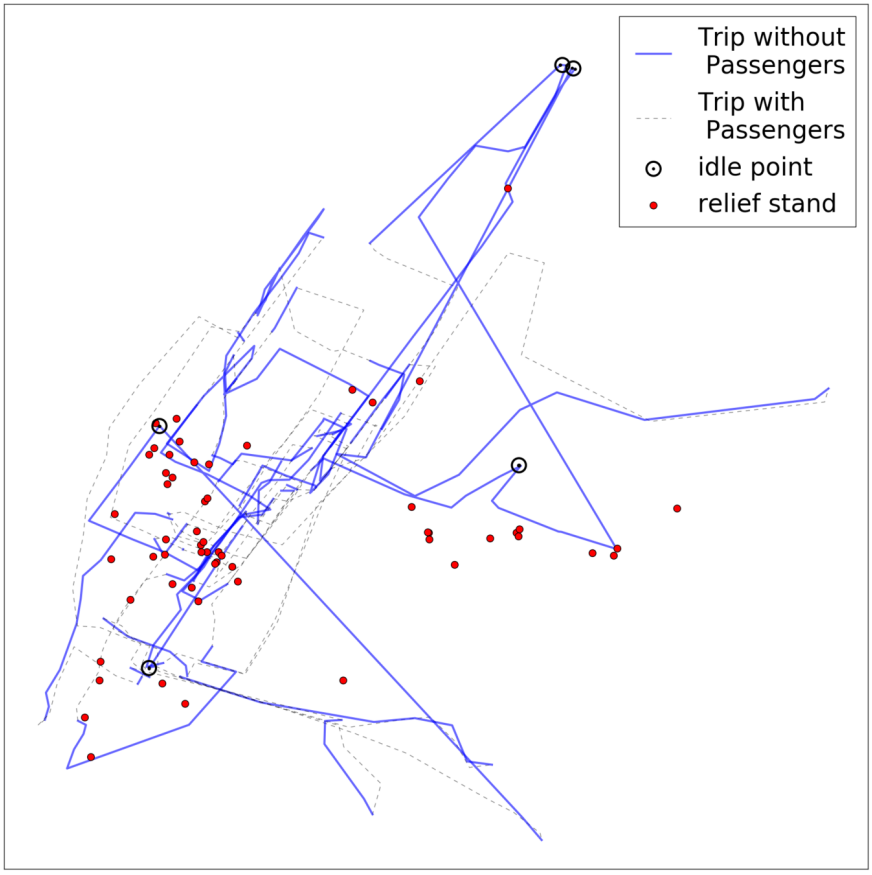

Figure 6: The one-day trajectory activity of a taxi is displayed above. Idle stays that are over 30 mins are detected and marked in the black circle. This plot helps in understanding the taxi driver behaviors as well as the relief stand usages.

In the figure below, the aggregated average daily drop-off counts in the year 2015 over hexagonal zones are displayed.

7.4

The analysis part is subdivided into two main parts.

- **Effectiveness of Existing Taxi/FHV Relief Stands**

The data processed through spark is used to analyze the yearly, monthly and weekly trends in the usage of the current relief stands. The time-series of the usage is analyzed by using 3-sigma thresholds to understand some of the events. Different statistics are calculated which are shown in the results section.

- **Suggesting locations for new Relief Stands**

There is no scientific definition of a good Taxi/FHV Relief Stands. All the models created for the project are based on discussions held with the client as well as some domain knowledge about working with the taxi drivers. The results of the models (suitable locations selected) are analyzed to understand their characteristics (violations, dropoffs, restrooms in that location). The results are also compared with the stands requests

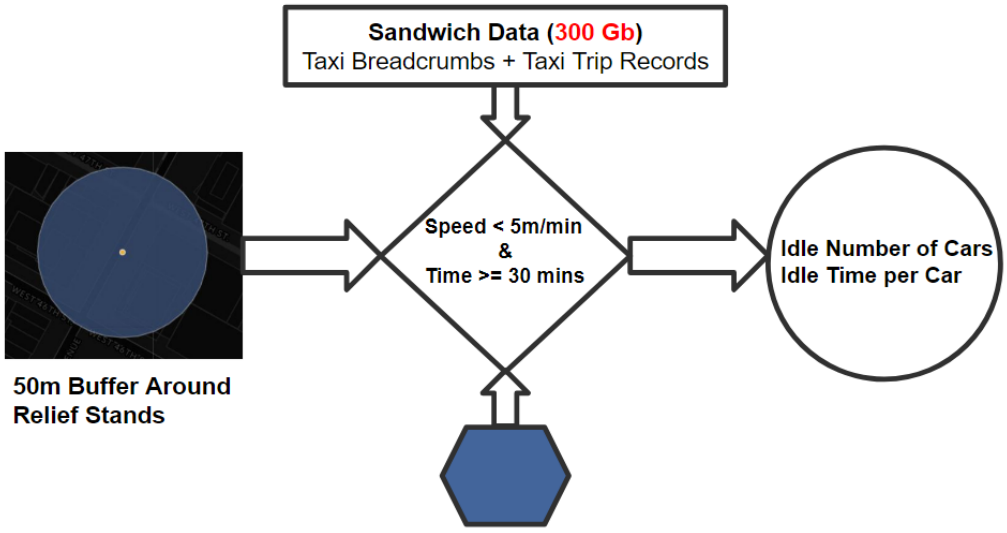

Figure 7: The flowchart of how the relief stand usage is being calculated.

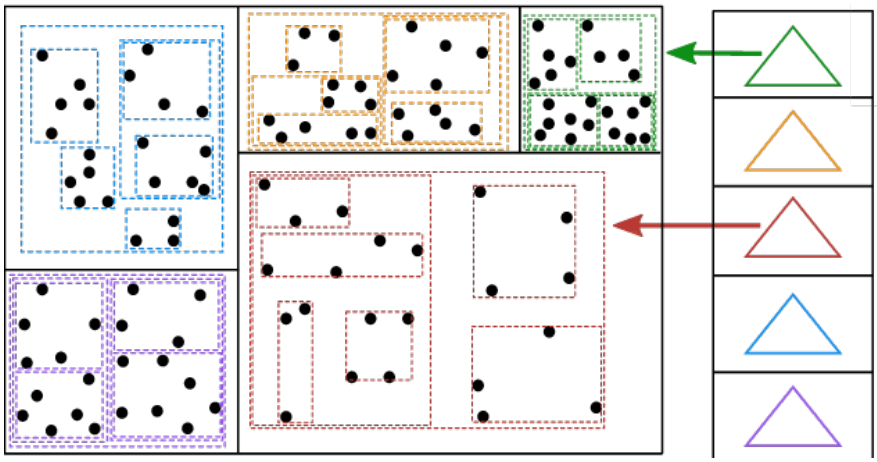

Figure 8: R-tree representation Image by Francois Anton (ResearchGate)

obtained from the requests dataset the TLC has provided. However, careful interpretation is recommended because the requests are from individual taxi drivers and it may be a random sample.

Different models are created and implemented to find suitable locations for new Taxi/Relief Stands. While creating models for such a problem where there is no definition of a good output, the problem is to be analyzed from different angles. Three models are implemented to find new locations.

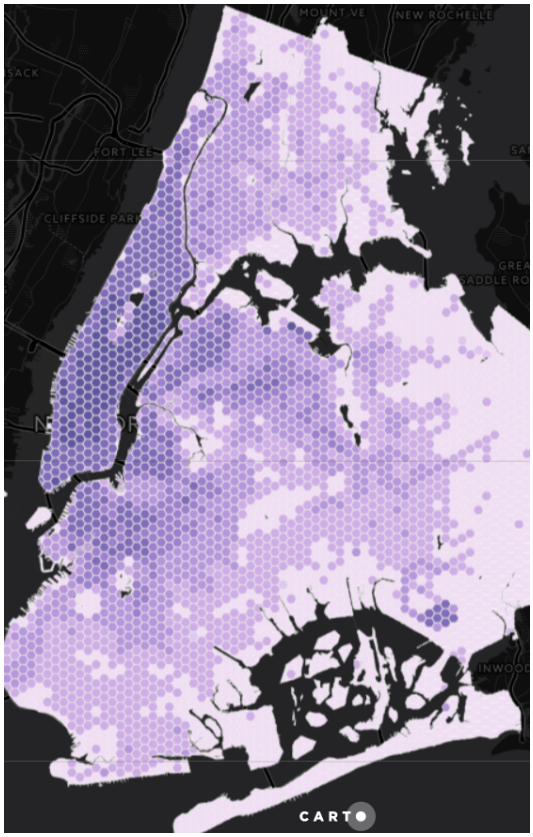

Figure 9: The choropleth map shows the aggregated average drop off counts per day over the year 2015.

7.5 Analytic models

Parks-Dropoffs Model
Rank Function
Mixed Integer Linear Programming

1. Parks-Dropoff Model

This is a simple model. Areas, where there are parks with restrooms, are good places to put a taxi relief stand as drivers can use the restrooms and relax. All the hexagons with parks (having restrooms) are selected and sorted descending by the drop-off count and the top hexagons are selected as suitable locations for new relief stands. The park denotes the need and the drop-off count denotes the demand. Since it is understood that drivers would prefer relief stands near high demand areas.

2. Rank Function

A rank function is applied in order to generate a "relief stand friendly" score for each hexagon. The rank function is defined as for each hexagon, the important features are learnt from current relief stands as constraints. The rank function can be written as:

$$rank_i = \sum_{j \in F} a_i f_{ij}, \forall i \in H$$

$$where \sum_{j \in F} a_j = 1$$

Here, f_{ij} denotes the score of feature j affecting hexagon i. The weights satisfy the normalized condition. Weight for each feature a_j $(j \in F)$ is carefully decided according to current relief stands usage results and experts' opinions. After consulting the TLC, higher weight is given to violations, the importance of the other features are in the following order: idle counts, drop off times, parks (with restrooms) and the last one is restaurants. 10 random number sets are generated, each set contains 5 random numbers to match each feature mentions above, using dirichlet random number generator to make sure that all the weights are added up to one. With these weights, the score for each hexagon in NYC is calculated and result is plotted. Finally, an averaged score of ten scores is obtained for each hexagon. Top 70 hexagons out of 4003 are shown as the potential areas that might be suitable for a relief stand. These hexagons are analysed for their characteristics.

3. Mixed-integer Linear Programming(MILP) model

This model is based on using the violation rate as an indication of the taxi drivers' parking demand within the hexagon. Thus the mixed-integer linear programming model that focuses on a set of hexagons H located in Manhattan is developed to optimally select regions for placing new taxi relief stands based on maximal covering the current parking demand. The model is also implemented for the complete NYC, since, most of the violations are concentrated in Manhattan as seen in figure 3, the main focus of the model is Manhattan.

In this model, the total number of new taxi relief stands is no more than R and each hexagon should at most have one new stand. Therefore a binary variables $y_i \in \{0, 1\}$, $\forall i \in H$ is used to decide if a new taxi relief stand is located in hexagon i. A new taxi relief stand helps to not only decrease the violation rate in its own hexagon but also decrease the violation rate in its neighbors. A hexagon i is covered with weight $\omega_0 \in [0, 1]$ if a new relief stand is located in hexagon i or weight $\omega_1 \in [0, 1]$ if a neighbor in N_i contains a new relief stand.

The reduction in violation rate in hexagon i is also proportional to the car usage rate. The usage rate of hexagon with relief stand is defined as the ratio of car usage and idle point count. Usage rate is learned from the hexagons which have relief stands.

A linear regression model is developed to predict the usage rate in the hexagons that has no relief stands using the current ones as training set and number of pick-ups, number of drop-offs, number of hotels, number of parks, number of cheap restaurants are used as features.

$$r = f(N_{pickup}, N_{dropoff}, N_{hotel}, N_{park}, N_{restaurant})$$

For a set of hexagon H, each hexagon i is assigned a value v_i , $i \in H$, counting the violation and $s_i, i \in H$, counting the idle points within this hexagon. $\overline{H}$ denotes a subset of H which already contains taxi relief stands. This Mixed-integer Linear Programming(MILP) model is written as:

$$max \sum_{i \in H} \frac{v_i r_i}{s_i} x_i$$

subject to:

$$\sum_{i \in H} y_i \leqslant R$$

$$x_i \leqslant \omega_0 y_i + \omega_1 \sum_{j \in N_i} y_j$$

$$x_i \leqslant \alpha s_i, \forall i \in H$$

$$y_i \in \{0, 1\}, \forall i \in H$$

In practice, the values set are $R = 20,\ \omega_0 = 0.8,\ \omega_1 = 0.2,\ \alpha = 0.1$.

The MILP model is solved using the "PULP" package available in Python and the 20 relief stands obtained as output are plotted and the resulting location's characteristics are analyzed.

8 V. Result

8.1 Relief Stand Usage Statistics

The breadcrumbs and the taxi ridership data is processed in Spark to get information about the usage of the taxi relief stands. The car usage is the average number of cars using the relief stands per day and the time usage is the average number of hours the relief stands is used per day. The statistics presented below are calculated from the datasets used in the project.

The average number of taxis/FHVs using the taxi relief stands per day is about 550 and the average time the taxis/FHVs spend at the relief stands per day is about 1150 hrs. A single taxi/FHV spends about 2hrs at the relief stand during the whole day. The data analyzed shows that 60% of the taxis/FHVs use the same 10% of the top used relief stands. The top and least used Relief Stands' lists are shown in Table 1 and Table 2. The Stands locations are shown on the map in figure 10.

Location	Time Usage (hrs)	Car Usage
36th St between Skillman Ave and 43rd Ave	274.369763	99.85205500000001
Park Avenue between E 135th St and E 138th St	170.228819	76.764384
Lexington Ave between E 28th St and E 27th St	77.613121	66.50137
E 28th St between Park Ave S and Lexington Ave	40.50862	34.123288
W 44th St between Eleventh Ave and Tenth Ave	68.851008	30.635616
Fourth Avenue between Third St and Sixth St	74.487917	28.268493
W 21st St between Sixth Ave and Seventh Ave	82.728021	27.205479
Lexington Ave between E 25th St and E 26th St	18.054627	14.019178
64th St between 34th Ave and 35th Ave	41.798145	11.671233
E23rd St between First Ave and Second Ave	17.879849	11.096154
43rd Avenue between 36th St and 37th St	26.884782	9.51
Third Avenue between E 14th St and E 15th St	22.326855	9.27

Table 1: Top Used Relief Stands (Daily Average)

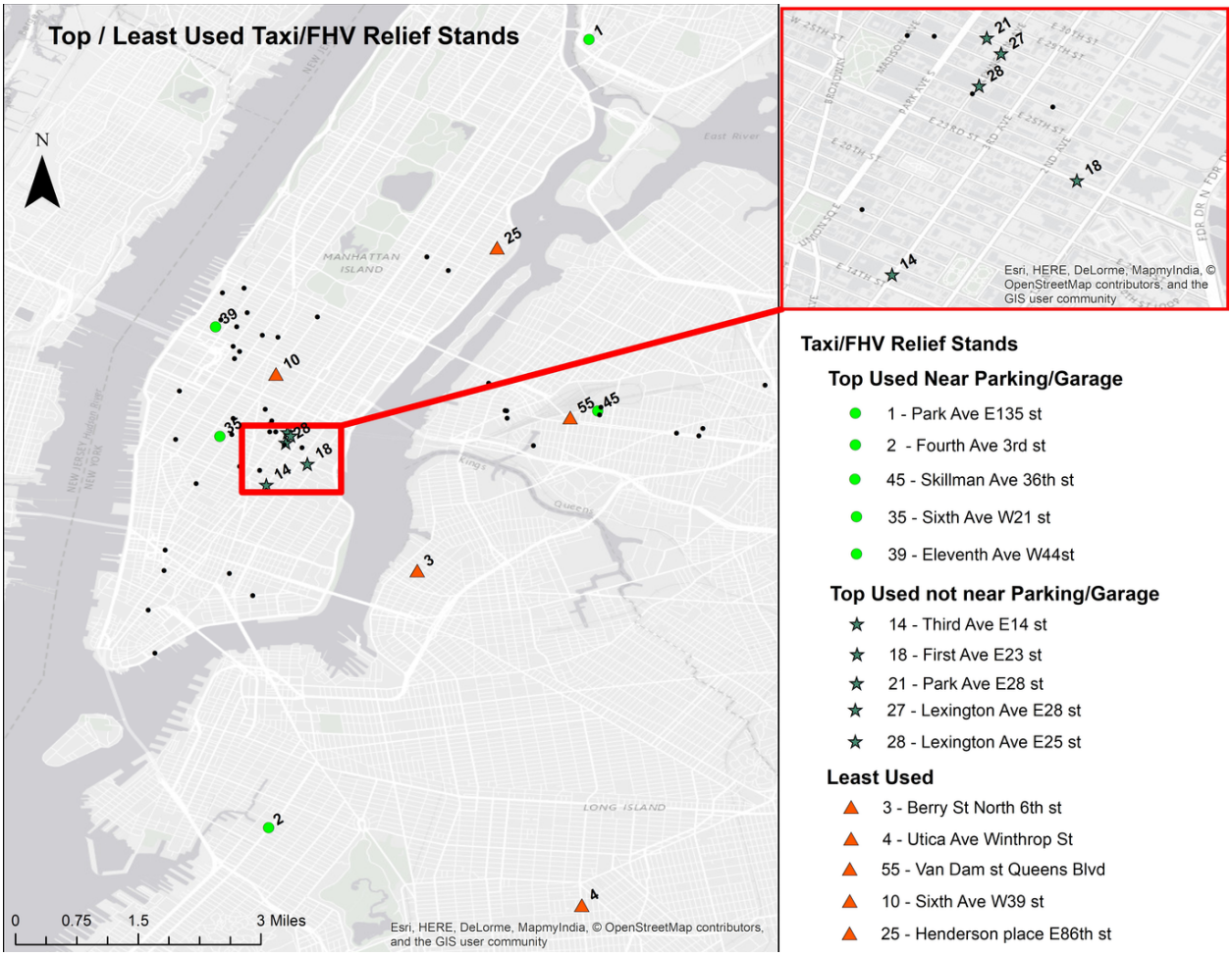

Figure 10: The map shows the top used relief stands in the red star which are mostly in Manhattan. The yellow star denotes the top used relief stands which are near garages/parking. The least used relief stands are shown by red triangles and are present mainly in the outer borough

Location	Time Usage (hrs)	Car Usage
Van Dam St between Queens Blvd and Skillman Ave	2.15	1.29
E 86th St between Henderson Place and York Ave	1.80	1.26
Sixth Avenue between W 39th St and W 40th St	2.10	1.21
North 6th Street between Berry St and Wythe St	1.46	1.20
Utica Avenue between Winthrop St and Rutland Blvd	1.64	1.17
W 41st St between Eighth Ave and Ninth Ave	2.21	1.13
45th Ave between 45th Ave Rdwy and 23rd St	4.26	1.12
E 77th St between First Ave and Second Ave	2.24	1.12
Pearson St between Jackson Ave and Dead End	3.32	1.12
Hunters Point Ave between 27th St and 30th St	1.85	1

Table 2: Least used relief stands (Daily Average)

Most used relief stands are defined by the high number of cars using the relief stands. However, the further analysis of the top used relief stands reveals that some of the top used relief stands are placed near taxi

garage/parking spots. These stands are supposed to have high usage. We also unveiled that some most-used reliefs stands, based on the idle points calculation, were located around those commercial car repair places outside of Manhattan. For example, the Relief Stand at Fourth Avenue location between Third St and Sixth St at Brooklyn, and the Bronx Relief Stand at Park Avenue between E 135th St and E 138th St.

Figure 11: A relief stand is present on West 44th St between tenth avenue and eleventh avenue near this taxi garage

Some of the top used relief stands are placed near areas with food options. Some examples are the relief stands on the Lexington Avenue

The car usage and the time usage are plotted below which gives an intuitive sense of the types of relief stands present in NYC. Stands near garages/parking show different behavior than others.

8.2 Trends in Usage of Taxi/FHV Relief Stands

The time-series of the usage of the *60 (The relief stands on the same street on opposite sides are considered same in our analysis, so the number is 60) relief stands is shown in below figures. Different patterns are observed in the usage patterns and also some events are detected from the time-series plots using the 3-sigma threshold.

The figure 14 shows that the time usage of the Taxi relief stands is more during the summer as compared to the winter.

The usage trends at the top 10 highly used relief stands are almost similar to the overall relief stand usage rate. The important thing to notice is that the car usage does not change much during the spring and summer, but the time usage in summer is higher than in spring. The car as well as the time usage decreases in winter.

Some of the top used relief stands are near garages/parking and some are not. When these relief stands are segregated, new patterns emerge in the usage. The usage of the relief stands near the garage/parking is more during the summer. This is in accordance with the overall usage of 60 stands. However, the relief

Figure 12: The taxi stand at Lexington Ave between 28th street and 29th street. This area has a lot of Indian restaurants. Eg. the one in the picture, Kailash Parbat

stands which are not near the garage/parking do not follow similar trend, their usage actually go down during summer.

Figure 17 makes the case more clear. Usage in winter is low. The usage in summer and spring differs by the type of relief stands.

The plot of the usage of 10 least used relief stands is shown in figure 18. The usage is really low at these stations throughout the year.

The usage data is also used to analyze the weekly pattern in the usage of relief stands. Figure 19 shows the weekly trend in the car usage of the top used relief stands. The number of cars visiting the relief stands near the garage/parking is more during the weekdays than on weekends. The vice versa is applicable for the other highly used relief stands.

The hexagons zones in which the highly used relief stands falls also show some peculiar characteristics. The features like number of violations, the number of drop-offs, number of restaurants, number of idle cars have values greater than 75th quantile in these hexagons.

Figure 13: The areas with a bigger circle are relief stands with high car usage (avg. no. of cars per day) and darker color represents the time for which a car is utilizing a relief stand. The relief stands with high car usage (larger circle) and high time/car usage (darker color) are mainly the relief stands near the garage/parking. The relief stands with high car usage but small time/car usage (lighter color) are top used relief stands near food joints (Eg. Lexington Ave)

9

9.1 Parks- Dropoffs Model

The output of this model is dependant on the number of drop-offs in the hexagonal zone and the availability of a park with restroom in that hexagon. The top 10 hexagonal zones from this model are shown in figure 20. The stand requests which the TLC has received is used to cross-verify the result. However, the stand requests should not be used as a proxy for good relief stands.

9.2 Rank Function

The output of the rank function depends on the values of the features and the corresponding weights. The top 70 hexagonal zones selected by the rank function model are shown below. The zones in Brooklyn, Queens, and the Bronx are selected because of the high number of idle points in those areas. Most of the drivers live there and that's why the idle count is high in those areas.

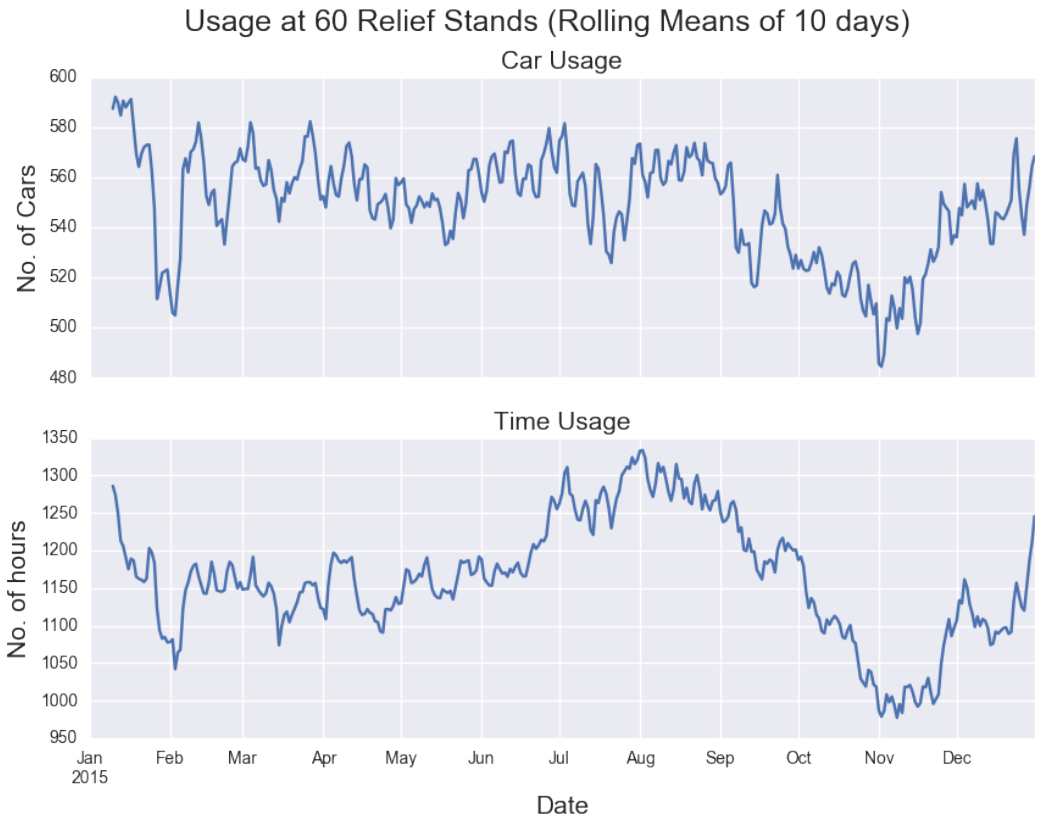

Figure 14: The time-series plot shows the time usage and the car usage of the 60 Relief Stands combined over the year 2015. The time usage increases during summer and decreases during winters. Similarly, the number of cars at the relief stands are less during the winter. The dip in usage around 27th January is because of the blizzard.

9.3 Mixed Integer Linear Programming Model

9.4

The MILP model provides a result which tries to reduce the violation rate in a zone and also considers the usage rate of the zone along with the influence of its neighbors. The output of the MILP model is calculated for the Manhattan. The MILP model becomes an NP hard problem when it runs for the whole of NYC. The result of the model is shown in figure 22

The top 20 hexagonal zones from the three models are shown in Table 3.

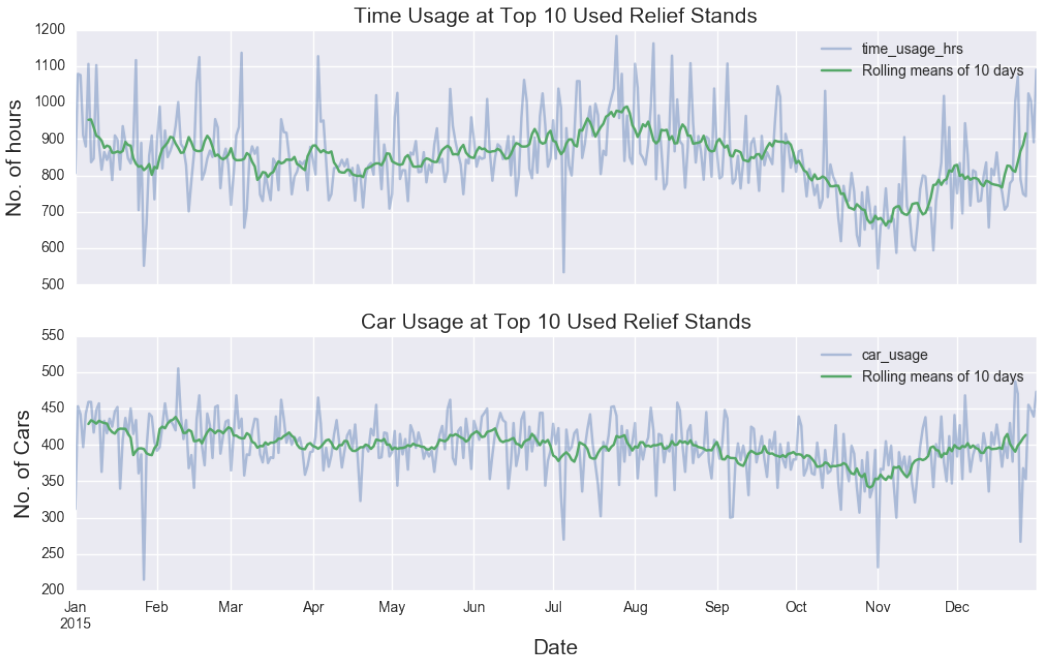

Figure 15: The time-series plot shows that the usage trend is similar to the trend of 60 relief stands combined. There are fluctuations in daily usage which are below 3-sigma thresholds. These events were as follows - 27th Jan was a blizzard. There is another dip on 4th of July (Independence Day). There was a marathon on 1st Nov. The last dip is around 25th December which is Christmas

Parks-Dropoffs	Rank	LP model
BB-43	BG-42	BG-33
BC-45	BD-38	BK-37
BB-41	BF-42	BL-32
BB-44	BE-39	BG-43
BG-44	BD-42	BN-38
BH-31	BH-40	AZ-54
BC-40	BE-40	BM-33
BN-32	BE-41	BI-38
BH-44	BG-37	BG-40
BJ-29	BI-38	BH-41
BM-23	BG-45	BL-34
BH-47	BH-39	BH-39
BC-49	BE-38	BG-41
BP-19	BI-39	BN-34
BO-34	BF-38	BC-37
BP-24	BB-43	BG-36
BP-26	BF-41	BF-37
BD-36	BH-41	BH-44
BP-31	BG-33	BB-51
BM-24	BD-41	BE-45

Table 3: Selected Top 20 Hexagonal zone IDs from the models

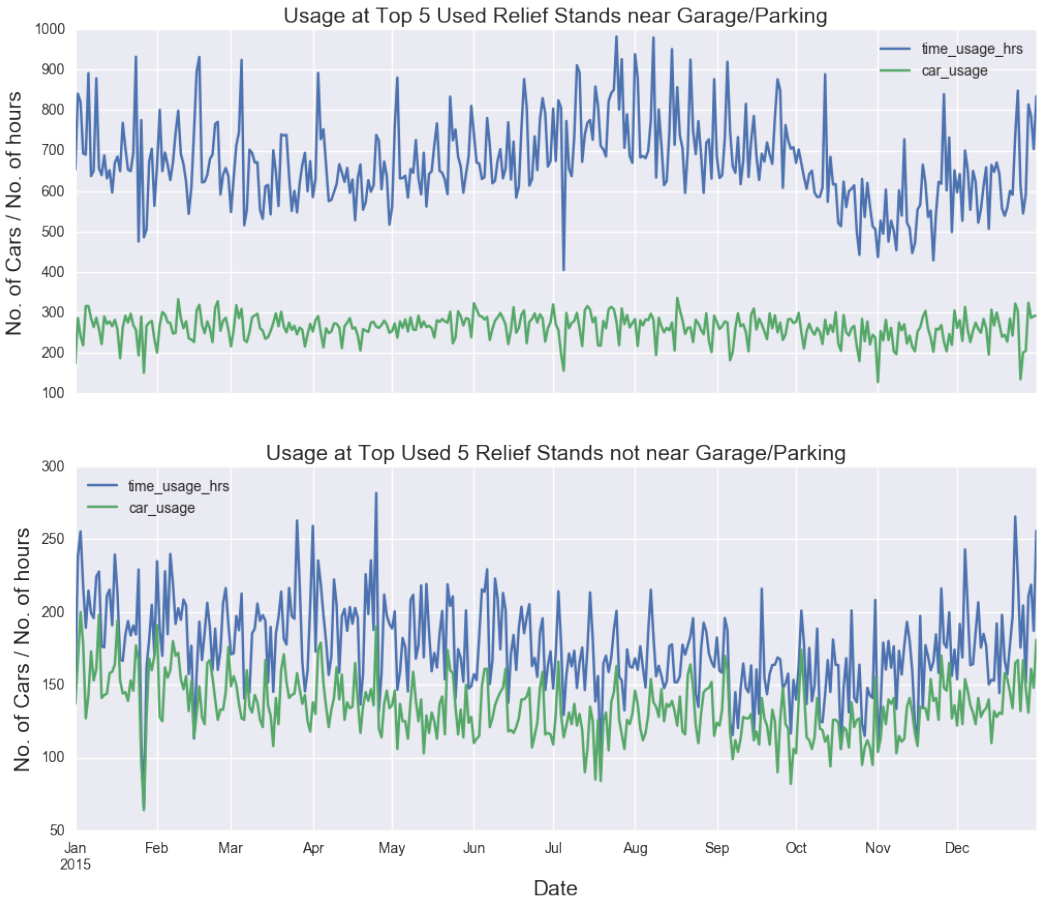

Figure 16: The comparative study of top used relief stands near garage/parking and the top used relief stands which are not near parking/garage shows that the usage patterns are different. The time spent at relief stands near garage/parking is more in summer than in spring. The time usage at relief stands not near garage/parking follows the opposite trend.

10

11 VI. Limitations

12

Every data analysis suffers from some kind of limitations due to lack of data, biases in data, biases in models. The project did have some limitations and biases which need to be kept in mind while interpreting the results.

12.0.1 Systematic Biases

Relatively bad GPS signals in remote suburban neighborhoods, where wifi or signal receptions are underdeveloped and also areas with tall buildings, would disturb the accuracy of breadcrumbs data, which is the data record of taxi location taken at every two minutes. This limitation was taken care while creating the algorithm by introducing a speed threshold.

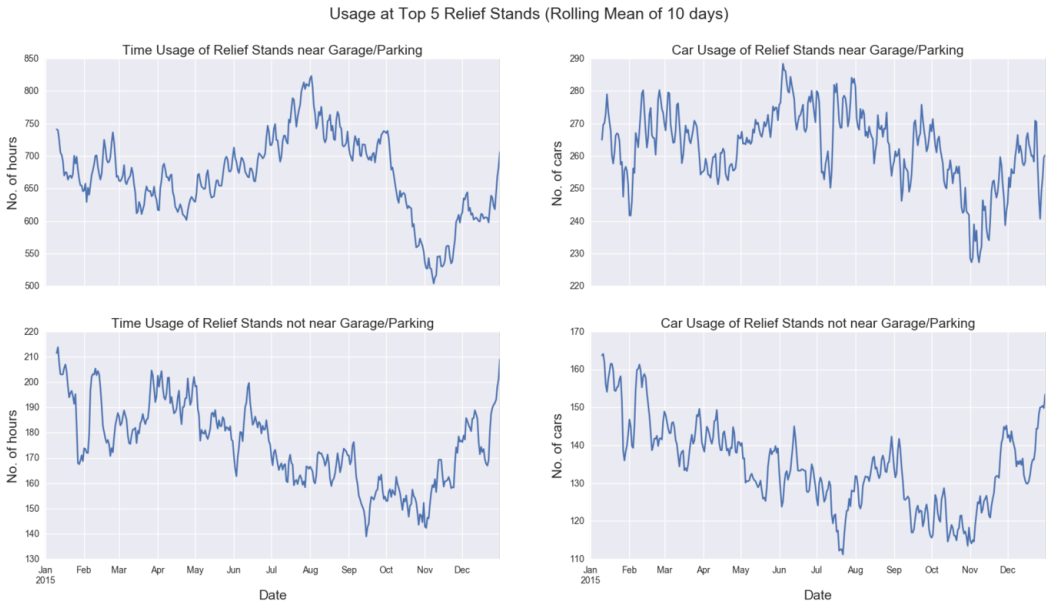

Figure 17: The plots compare the time usage and car usage at the top used relief stands near garage/parking and other relief stands. The taxi/fhv driver prefer using the relief stands near garage/parking during the summer.

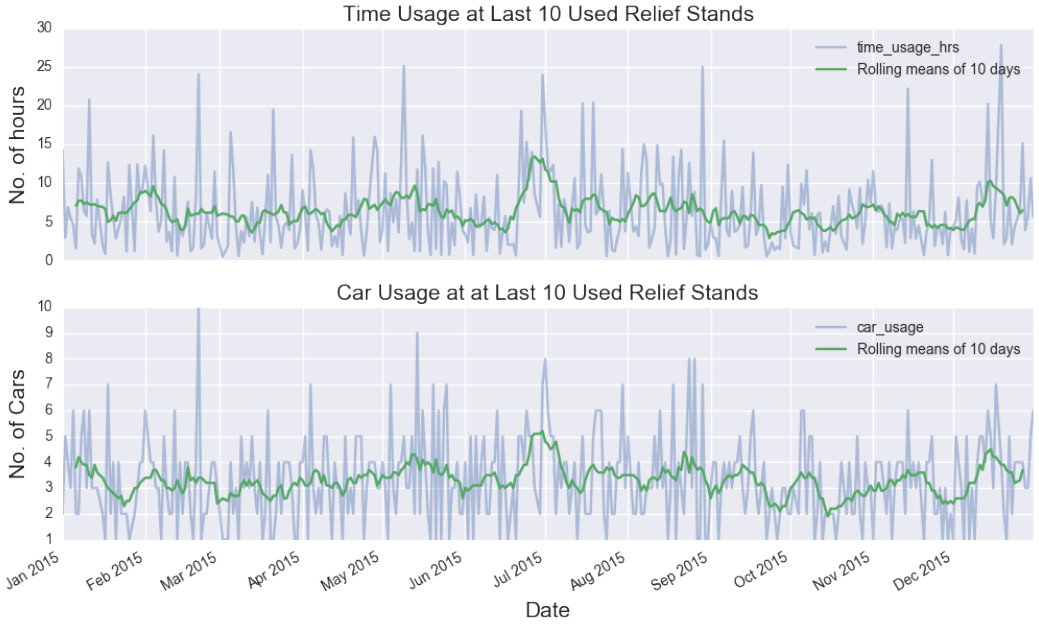

Figure 18: Time-series plot for usage shows that the usage is really low throughout the year and there are no significant changes.

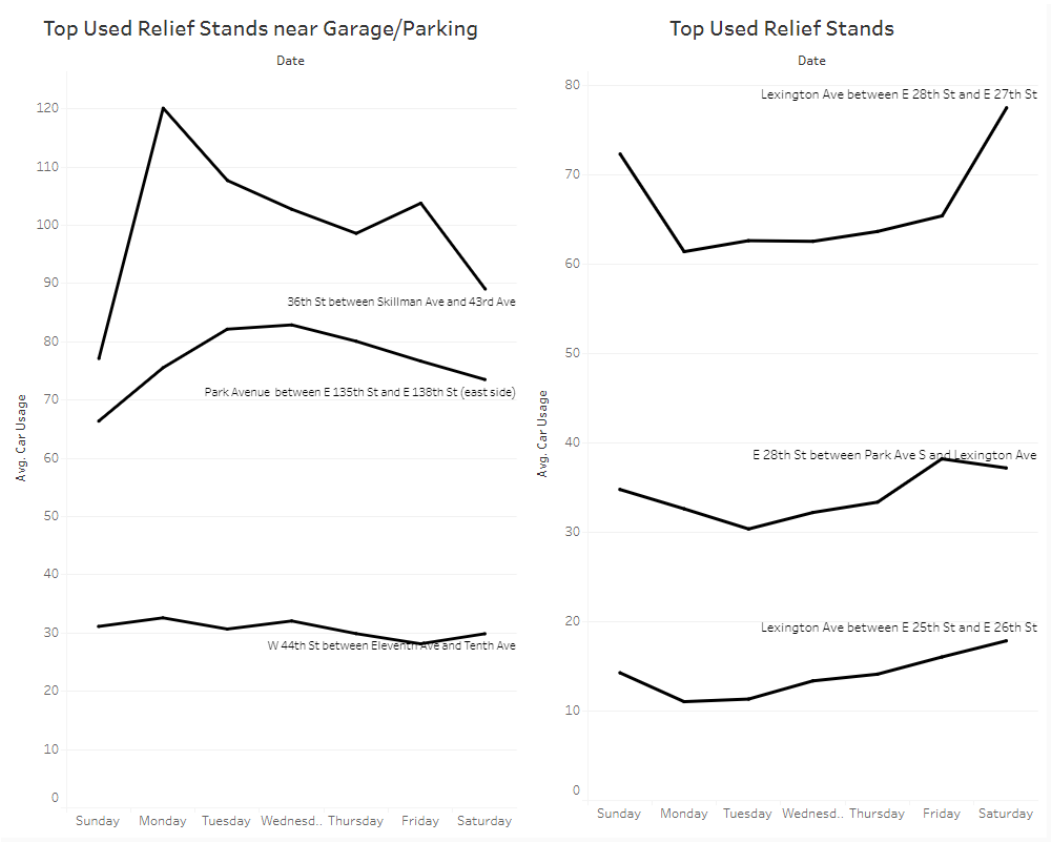

Figure 19: The weekly plot of car usage of top used relief stands shows that the taxi drivers visit the relief stands which are not near the garage more during the weekends than on weekdays.

12.0.2 Missing 2016 - 2017 up-to-date data

The project uses yellow and green taxi data from 2015 to stay consistent with the 2015 Breadcrumbs data we have in the CUSP data hub. Currently, the number of yellow and green taxi trips, as well as the number of these taxis, may have changed because of FHVs like Uber, Lyft, Via, etc. The article from a well-known analytics website FiveThirtyEight points out that "Throughout Manhattan, riders have shifted from taxis to Ubers, perhaps attracted to features Uber promotes as advantages: newer cars, no need to hail, driver ratings and no tipping." (Reuben Fischer-Baum & Carl Bialik, 2015). It is possible that the data being used is not an exact representation of the current situation.

12.0.3 Discrepancy on Relief Stand Data

Although the taxi relief stands data was listed on DoT's website, the problem of the discrepancy between various data sets has grown larger. The relief stands are not updated on the website, which offers no clue for drivers. There are 153 Taxi relief stands as per the Parking Sign data set. These differences among the taxi relief stand datasets have not only limited the ability to discover more characteristics of the relief stands, but also the validation process of the model.

12.0.4 Missing FHV data

As a matter of fact, the available taxi datasets do not include data about the FHVs trip records. The usage statistics maybe be biased due to the intrinsic difference of the characteristics of traditional taxi and FHV.

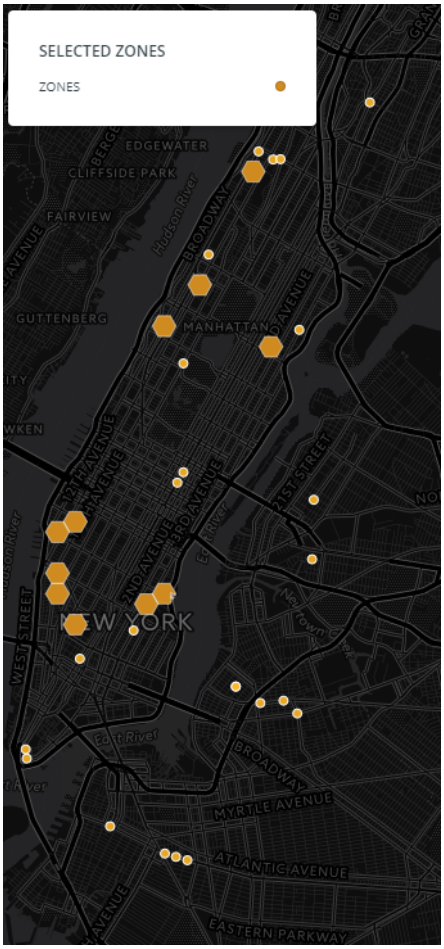

Figure 20: 5The top 10 selected zones are shown. The yellow dots are the stands requests the TLC has received. Some of the selected zones are near the requested area.

Namely, FHV (Uber, Lyft) are e-hailed car service providers, as they can be idle and receive customer orders by electronic devices. However, regular taxi (especially Yellow) has to roam on the streets to pick up random passengers. It is necessary to be aware of the fact that the relief stand usage data does not consider the FHVs usage at all. Undoubtedly, the usage statistics would be improved if the FHVs usage is included. The model created to recommend new taxi relief stands suffers from the lack of FHVs data

12.0.5 Limitation on other Features

What kind of restaurants should be taken into consideration is another tough question. Besides the various habitual behavioral patterns from a certain culture, the special food preference of different taxi driver groups also brings the attention of unfairness or biases during feature selection analysis. For this analysis, the restaurants that have public restrooms may not be included in the data set. The food truck data is missing in the analysis.

The project does have some limitations and biases present because of missing data, misrepresented data, technical biases which are known and corrective steps are taken to resolve wherever it is possible. The client is also aware of these limitations.

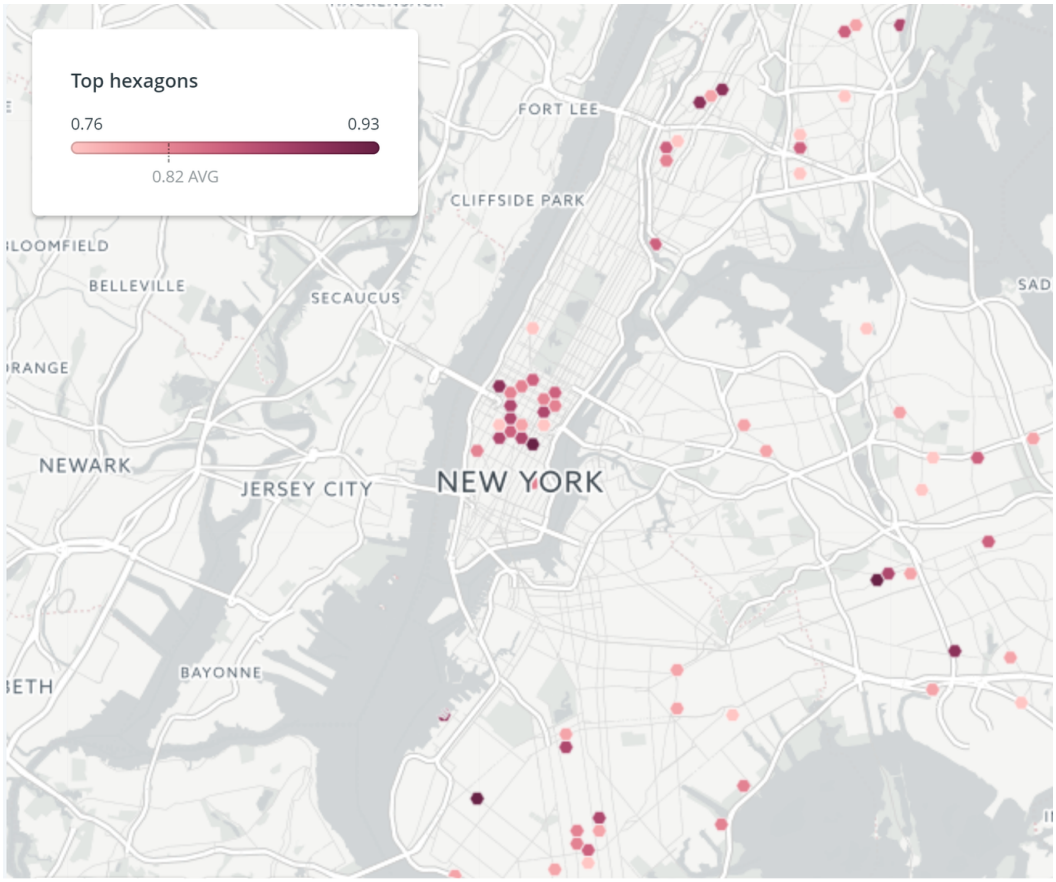

Figure 21: Top 70 hexagons obtained by score function, the darker the color is, the higher the scores are, scores are between 0.76 and 0.93, in Manhattan area, the scores are higher under central park because of higher park violations, higher drop off counts and many restaurants, some points in Brooklyn are high for the idle points and violations.

13 VII. Conclusion/Future Study

Benefited from the Hadoop system, more than 300 Gbs of data is successfully streamed through the CUSP cluster. The overall output not only offers valuable statistics on the taxi relief stand usage but also provides a decision support system for allocating new Taxi/FHV Relief Stands in entire New York City. The work with TLC has also offered a great understanding in the procedure of policy making as well as the importance of data privacy handling. Even though the data processing is the heavy-lifting part of this project, the Mixed Integer Linear Programming model developed should be carefully analyzed in order to offer the best stage for the data to speak on its own. The number of potential taxi relief stand recommended through two of the models which based on different feature importance, mainly focus on the parking ticket reduction as this will benefit the drivers by providing relief stands as well as reduced tickets. The further validation of the results is needed in order to better optimize the model. The missing FHV data can be used for optimizing the model, as well as the accurate information on the existing relief stands can help in improving the model.

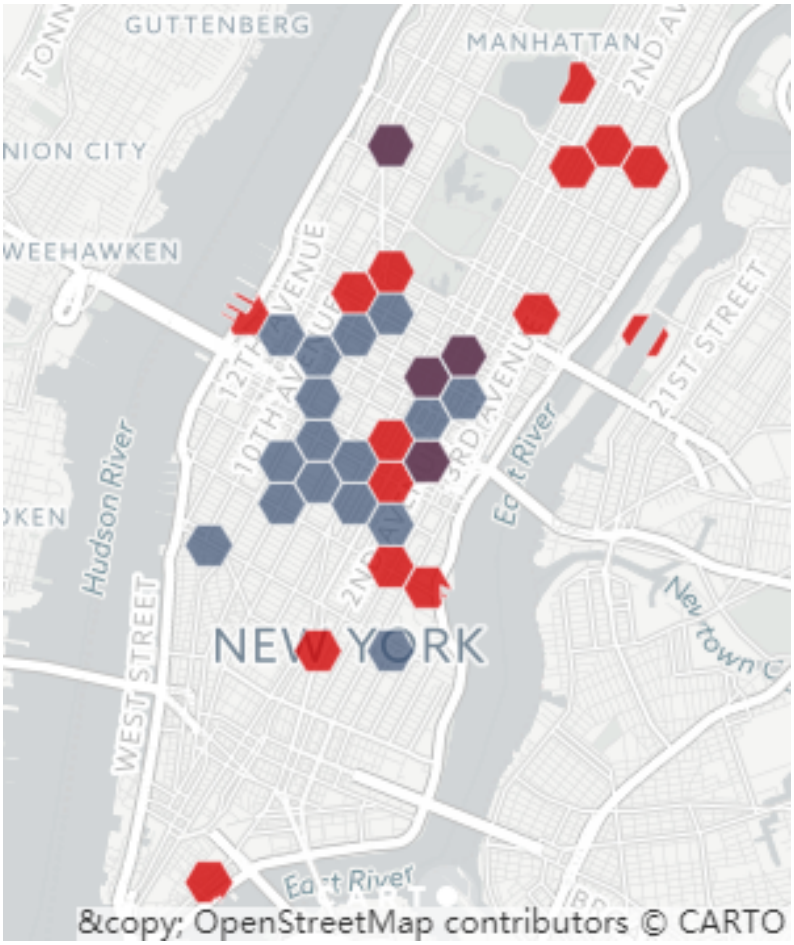

Figure 22: Hexagons with red color are the result of the LP model, and blue ones are from the top 20 by rank function model. Purple ones are the overlapping hexagons. Most of the hexagons from the LP model are located where parking violations are high and cheap restaurants are available nearby. In each of the red ones, one relief stand is recommended.

Collective opinions of drivers should be considered for future study. The relief stand usage depends on various factors, the relationship between these factors and the relief stand usage may not be linear, however, considering the situation, the use of linear model was deemed right by us it was more interpretable. However, in the future more complex model can be implemented to understand the relationship but being watchful while interpreting the results.

14

VIII. References

Johannes Asamer, Martin Reinthaler, Mario Ruthmair, Markus Straub, Jakob Puchinger. Optimizing charging station locations for urban taxi providers. Transportation Research Part A: Policy and Practice, 2016, 85, pp.Pages 233-246

Strimas-Mackey, Matt.(2016) ***"Comparing Benefits." Fishnets and Honeycomb: Square vs. Hexagonal Spatial Grids.***

MRCagney Pty Ltd, Brisbane (2012), ***"Taxi Rank Master Plan for Melbourne".***

NYC Taxi & Limousine Commission (2016). ***The TLC Factbook.***

Reuben Fischer-Baum & Carl Bialik (2015). ***Uber Is Taking Millions Of Manhattan Rides Away From Taxis.***

15 IX. Contributions:

"The strength of the team is each individual member. The strength of each member is the team."

– Phil Jackson

Cheng Hou(C.H), **Le Xu(L.X), Vishwajeet Shelar(V.S) and Yao Wang(Y.W)** worked together for the successful completion of the project.

C.H and **Y.W** did the big data processing in Hadoop and Spark. They also applied R-tree in the algorithm. **C.H** and **Y.W** worked together to write and debug the map reduce scripts for Hadoop streaming over both Breadcrumbs data and taxi trip data. **Y.W** and **L.X** developed the idle points counting algorithm. **C.H** and **Y.W** developed the models. **C.H** piloted the formula of Mixed Integer Linear Programming model and **Y.W** designed the Rank Function. **C.H** and **Y.W** also managed the LaTex writing of the final report.

V.S and **L.X** were in charge of the team building and shared responsibilities as the team administrator. **L.X** and **V.S** also took part in the model building process and assisted **C.H** and **Y.W** by preparing the necessary data for the models. **V.S** and **L.X** also prepared the statistics for the relief stands and co-authored the peer reviews. **V.S** handled most of the visualization and created the slides for the final presentations. **V.S** also oversaw the munging of the taxi relief stands, parking ticket data sets and studied each relief stands over google map. **V.S** also geocoded the parking tickets, relief stands, public park restrooms and relief stand requests. **V.S** developed the park and drop-off model. **L.X** created the Gantt charts and Initial exploratory analysis for the idle points counting and was also involved in data cleaning, data wrangling, and feature mining of Restaurants, Hotel, Parking Violation datasets. **L.X** also took over most parts of the report writing and assisted in presentation preparation.

16 APPENDIX

The shapefile for hexagonal grids is uploaded in the data folder.

All the codes created for the project are available here - **Link**