Traffic Flow and Density Analysis of NYC TLC Taxi Data

by

Zui Chen

An honors thesis submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Science
(Honors Statistics)
at the University of Michigan
2018

Supervisor: Dr. Long Nguyen

GSI: Aritra Guha

ACKNOWLEDGMENTS

It is imperative that I thank my supervisor, Dr. Long Nguyen, and my GSI, Ph.D. Candidate Aritra Guha, from the Department of Statistics for their advice and instructions. I also must thank Jiahui Ji for her contributions to my work.

TABLE OF CONTENTS

Chapter

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

NYC TLC New York City Taxi and Limousine Commission

DBSCAN Density-based spatial clustering of applications with noise

OPTICS Ordering points to identify the clustering structure

ABSTRACT

Traffic Flow and Density Analysis of NYC TLC Taxi Data

by

Zui Chen

Supervisor: Dr. Long Nguyen

This paper analyzes the trip record data collected and provided to the New York City Taxi and Limousine Commission (NYC TLC). The analysis is mainly focused on the geographical location and movements of taxi trips. I analyzed the 2015 and 2016 green taxi (Boro taxi) data. I found that features such as traffic flow, orientation and range were of particular importance for the taxi trips.

CHAPTER 1

Introduction

Everyone whoever goes out into the streets is involved in traffic. In the urban cities, especially in metropolises, various issues occur correspondingly. For example, the traffic congestion has not only lowered the efficiency of the commute speed but also resulted in huge amount of loss in money. Consequently, it is important to study and understand the traffic patterns so that we could develop better urban plans and transport designs.

The most common form of urban traffic is the ground transportation. A large number of people drive in their own vehicles while the rest choose to take public transportation, such as buses. In the United States, every 1000 people have 910 motor vehicles (1). Besides, taking a taxi is another popular choice among many people when they do not have access to private vehicles or general public transportation. In a metropolitan like New York City, there are over 1,400,000 trips by Boro taxis in a typical month. Given such large amount of data, it would be interesting to study the patterns of urban traffic using the taxi trip record data.

Traffic flow analysis is the study of the movement of individual drivers and vehicles between two points and the interactions they make with one another. Unfortunately, studying traffic flow is difficult because driver behavior cannot be predicted with absolute certainty. However, drivers tend to behave within a reasonably consistent range. Therefore, traffic streams tend to have some reasonable consistency and can be roughly represented mathematically. To better represent traffic flow, relationships have been established between the three main characteristics: flow, density, and velocity. These relationships help in planning, design, and operations of roadway facilities (2). I am interested in the flow and density the most. In this paper, I analyzed the flow and density of NYC green taxi trips using the data provided by the NYC Taxi and Limousine Commission. My study is motivated by the following questions:

1. Where are the centers of different commuting areas?

2. How to predict the commuting algorithm based on different time of the day?

These two questions will be helpful to understand the commuting patterns by taxi in the New York City throughout a typical day and will be useful to not only providing taxi arrangement for the taxi company but also facility development for the government, as well as a reference for the passengers when planning their trips.

1.1 Background

The study of traffic flow and density are highly correlated to a wide range of social issues such as public health, economic growth, and city development (3). Researchers have been studied traffic flow data to predict the traffic trends. In the study on the traffic congestion and relationships of traffic flow, Bernard used historical data from the Georgia Department of Transportation to discover trends in traffic volume data. By using the available data from the year 2003, she attempted modeling time-based trends in the data using smoothing regression functions (4). Lu etal. studied the real-time traffic flow state identification and prediction based on big data-driven theory. They built a bi-level optimization model for regional traffic flow correlation analysis and to predict traffic flow parameters based on temporal-spatial-historical correlation (5).

Meanwhile, the study of traffic density could result in better understanding of congestion and lead to wider awareness in the public and further improvement on the problem. For instance, the INRIX Global Traffic Scorecard provides an evaluation of urban travel, traffic health and vibrancy for over 1,000 cities around the world (6). It could also be considered as an indicator for population density. In previous studies, Deville etal. used mobile phone call records to make spatially and temporarily explicit estimations of population densities (7); Khodabandelou etal. proposed a new approach to infer population densities at urban scales, based on aggregated mobile network traffic metadata (8). Taxi trip records, which has clearer information on the geographical locations, can be used to measure the popularity of areas in the city.

1.2 NYC TLC Taxi Trip Data

The taxi trip records include fields capturing pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types, and driver-reported passenger counts (9). Each data set display these information for a single month. In my research, the target feature is the locations. Meanwhile, I have also taken the times into account. The Travel Time Data Collection Handbook suggests that the four time elements for consideration can be month, day of week, day type, and time of day (10)). In my analysis, I chose to study the pickup and drop-off locations based on the time of day.

VARIABLES	DESCRIPTION
`pu.time`	The time when the meter was engaged.
`do.time`	The time when the meter was disengaged.
`Pickup_longitude`	Longitude where the meter was engaged.
`Pickup_latitude`	Latitude where the meter was engaged.
`Dropoff_longitude`	Longitude where the meter was timed off.
`Dropoff_ latitude`	Latitude where the meter was timed off.
`Trip_distance`	The elapsed trip distance in miles.
`Trip_timing`	The time duration of the trip in seconds.

Table 1.1: List of variables used in analysis

Noted that, due to privacy issues, trip records since July 2016 use Location ID, each corresponding to a bounded zone of neighborhood, instead of longitudes and latitudes. Therefore, my analysis on geographical coordinates (lon, lat) is based on date before July 2016. According to preliminary exploration of the data, the distributions of the trip locations and time are similar in each month. Therefore, in order to improve computation efficiency, I chose the data from September 2015 for the analysis. A part of my research on traffic flow included an exploration on the data with spatial taxi zones. For this part, I chose the data from September 2016 to learn the patterns.

1.3 Data Preprocessing

The original data contain geographical as well as time information for each pickup and drop-off. In order to study the commuting algorithms throughout a day, I parsed all the data from 2016 into 24 pick-up hour periods, each starting from 0:00 to 23:00. Each location is indicated by a pair of coordinates of longitude and latitude in five decimal places. I rounded these coordinates to four decimal places as if the map of New York City being parsed into small 10-4 by 10-4-unit degree squares.

1.4 Exploratory Analysis

Figure 1.1 on trip distance and Figure 1.2 on trip duration indicate that both variables follow a gamma distribution with the existence of some outliers. This provides the support for the assumptions in later analysis.

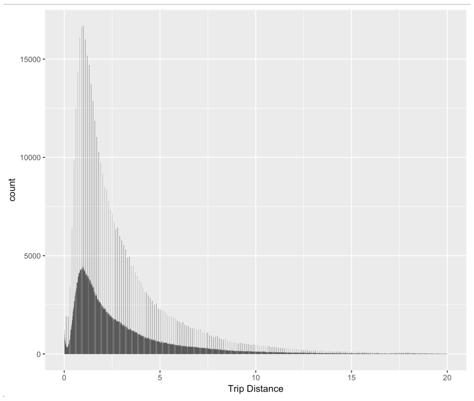

Figure 1.1: Distribution of Trip Distance

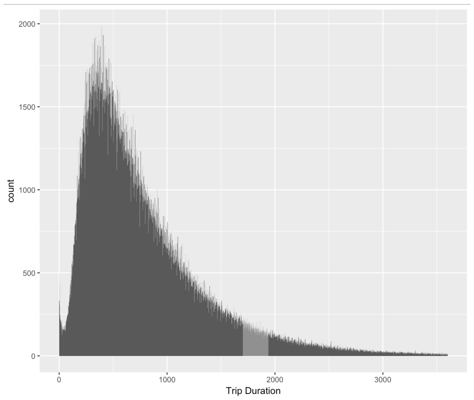

Figure 1.2: Distribution of Trip Duration

Figure 1.3 shows a series of pickup and drop-off locations within 12 different hour periods of a day. It clearly shows a change of the range of pickups and drop-offs along with the change of time.

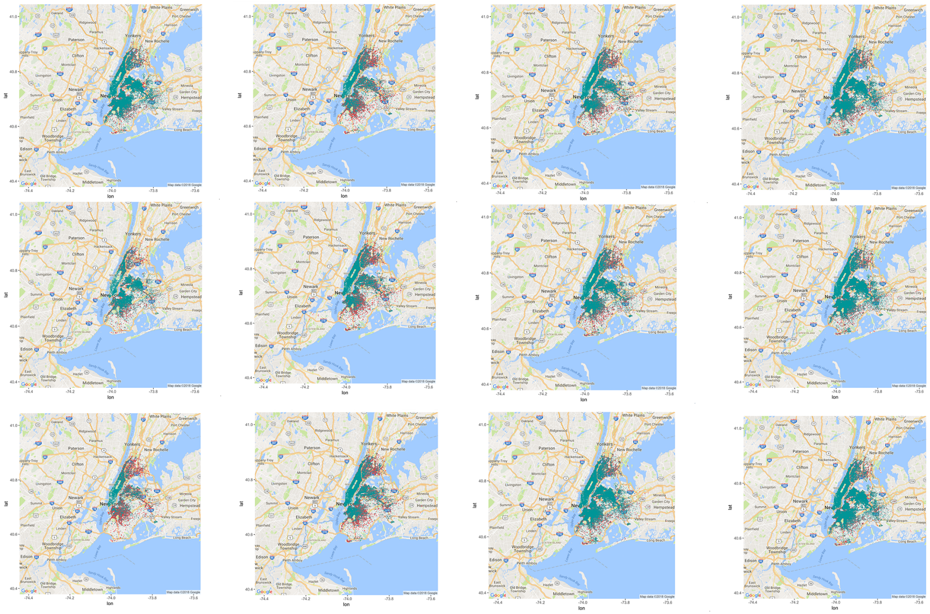

Figure 1.3: Pickups (red) and drop-offs (green) of trips at pickup hour 00:00, 03:00, 06:00, 07:00, 08:00, 09:00, 11:00, 14:00, 17:00, 18:00, 19:00, 20:00 (top to bottom and left to right)

CHAPTER 2

Traffic Flow

This chapter summarizes the work I did from last semester on the traffic flow. I used the data from September 2016, which has information about the pickup and drop-off zones so that I could see the flow of taxi trips between the neighborhoods.

2.1 Spatial Taxi Zones

Figure 2.1 on the left displays the divisions of all 263 spatial taxi zones. Figure 2.2 on the right showed the top 10 flow directions in September 2016. Each arrow represent the flow between two different zones while each circle indicates the trips taking place within the same zone.

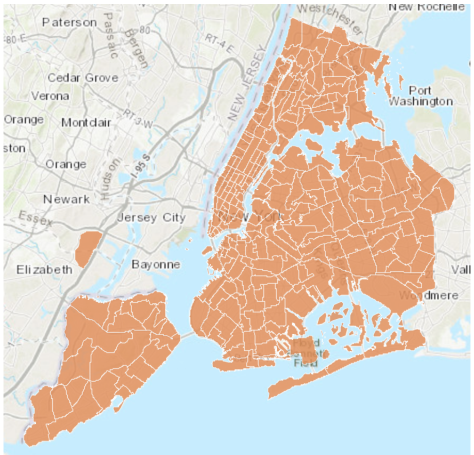

Figure 2.1: Green Taxi Zones

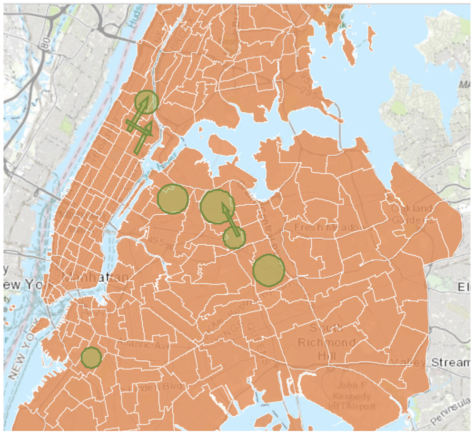

Figure 2.2: Top 10 Flow Directions in September 2016

2.2 Volume of Traffic Flow

I used Sankey diagrams to present the connections between starting locations and destinations. It could explicitly showing the largest flow during a certain time of day. In each Sankey diagram, nodes on both sides represent the taxi zones. Pickup locations are on the left and drop-off locations are on the right. The linking edges represent the trips, directed from the left to the right. The weight of each edge equals to the number of trips between the two zones. It is the indicator of the volume of the traffic flow between the two location zones.

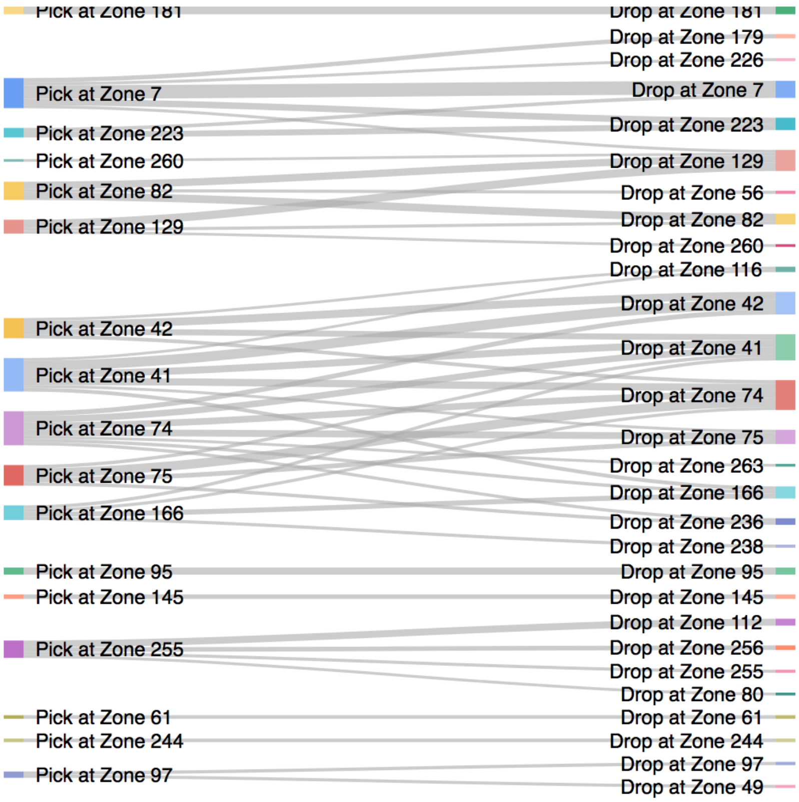

Figure 2.3: Top 50 Flow Directions in September 2016

From Figure 2.3, it is clear that a considerable number of trips happen within the same zone. Therefore, it is intuitive that there exist ”traffic circles”, in which the likelihood of someone taking a taxi is higher. This leads to my further study of the density of traffic based on the taxi trip records. I would also to find out if there exists any relationship between the time of the day and the location of the trips.

CHAPTER 3

Traffic Density

Traffic density is a fundamental macroscopic characteristic of traffic flow, and is used in assessing traffic performance from the point of view of users and system operators. It is also employed as the primary control variable in freeway control and surveillance systems (11). The difficulty in measuring density inhibited its general use until the early 1960s, when presence-type detectors were introduced (12). Density is also an important measure of the quality of traffic flow, as it is a measure of the proximity of other vehicles, a factor which influences freedom to maneuver and the psychological comfort of drivers (13).

3.1 Bouding Traffic Circle

Regardless of picking up or dropping off, it is possible to divide the location points into bounded traffic circles. In each circle the radius maximize the the average increasing ratio of the location points correspondingly. The traffic circles can provide information on the density of pickup and drop-off locations. In these manually divided traffic circles, passengers travel within the boundaries.

The generation of bounded traffic circles can be summarized into the following steps:

i Find an initial point, e.g. (-73.99488,40.69722), set as the center C

ii Initialize a value for the radius: $r = 0$ to the enumerate environment.

iii Calculate the number of location points within the circle centered at C with radius r as k_1

iv Increase r by a constant stp, e.g. 0.0006

v Repeat iv. n times until $k_n < (1 + stp)^2 * k_n - 1$

vi Take the $(n - 1)$th radius r and plot the circle C

This division largely depends on the functionality of each district. The more multi-functional the area is, it is more likely to become a circle district. If an area contains working, commercial, and entertainment facilities, then it is usually more compacted and can hold more within-the-bound taxi trips. By finding out the locations with various facilities, I was able to manually select the centers of the traffic circles. Complementing with K-means clustering, I obtained the circles in Figure 1 for a random selection of 10000 observations from data 8:00 am to 9:00 am.

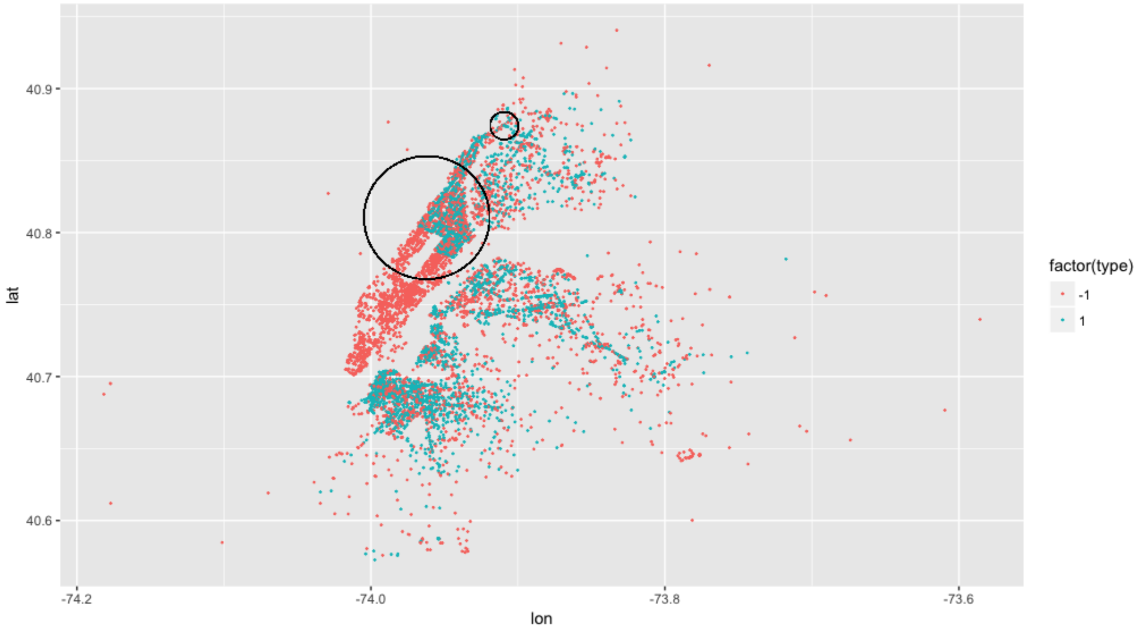

Figure 3.1: Boundaries for traffic circles 8:00 am to 9:00 am

In the figure above, the two largest circles are located around Upper Manhattan and Inwood. Upper Manhattan has been historically the town center where people can work and have entertainment. Therefore, it is a central area within which the taxi trips tend to be made inside the boundary. The typical volume of people is large enough to provide the taxi driver customers within the boundary. Inwood has become a new entertainment center in recent years, where a lot of new restaurants and lounges have been built. Besides, there are several parks in the district.

The traffic circle division is also relevant to the public transportation availability as well as the construction of the infrastructure. Considering the distance, convenience, and price, people's choices of whether taking a taxi or not vary in different cases. Consequently, the radius of each circle can vary a lot. Besides, the street layout can influence the radius.

The division of traffic circles based on the range of trips made within the boundaries could be a alternate approach to learn about traffic flow. It differentiate the in-bound and out-bound traffic flow by providing the critical point of the two directions.

However, this method of analyzing traffic density is really time consuming and requires a considerably detailed perception of the data. The choice of initial data point also affects the result largely. Moreover, it is not capable of bounding traffic circles precisely because it only results in boundaries in the shape of closed circles. Therefore, some more robust way of detecting the traffic circles are needed.

3.2 Density-based Clustering

From the previous plots in Figure 1.3, I have learned that the range of pickup and drop-off locations vary from hour to hour. This change results in different traffic centers, in which the trips are more ”dense”. Studying of these traffic centers will provide a better understanding of the popular destinations at different time of the day. From this insight, not only can we find out the popular places people heading to at different time, we can also learn the patterns to better facilitate the urban traffic planning, including the design of ground transportation routes and the promote of carpools or ride shares.

In order to find these ”traffic centers”, I used unsupervised clustering to detect the cluster in the drop-off locations. Instead of using K-means, I implemented density-based clustering algorithms, which do not require a prior specification of number of clusters and are able to identify outlier points while clustering. These algorithms can find clusters in arbitrary sizes and shapes. Partitioning methods such as K-means clustering is suitable for finding spherical-shaped clusters or convex clusters. They work well only for compact and well separated clusters. Moreover, they are also severely affected by the presence of noise and outliers in the data. Considering the size of data, an average of 61353 trips per hour, and the irregularity of the shapes formed by the location points, density-based clustering algorithms are more suitable for the taxi data.

3.2.1 DBSCAN

DBSCAN algorithm, is a density-based clustering algorithm, first introduced by Ester et al. in 1996, which can be used to identify clusters of any shape in a data set containing noise and outliers (14). Unlike K-Means, DBSCAN does not require the number of clusters as a parameter, but rather infers the number of clusters based on the data. The formed clusters are dense regions in the data space, separated by regions of lower density of points. The DBSCAN algorithm is based on this intuitive notion of clusters and noise. The key idea is that for each point of a cluster, the neighborhood of a given radius ϵ has to contain at least a minimum number of points denoted as $minPts$. The pair of ϵ and $minPts$ can be chosen from the observation of K-Nearest Neighbors distance. For a chosen $minPts$, ϵ will be the $minPts$-NN distance at the "elbow" of the KNN distance plot.

We can recap the DBSCAN algorithm into the following steps:

i For each undefined point P in the dataset S, calculate the number of points N within its neighborhood of distance ϵ.

ii If $N \geq minPts$, then these N points and P form a cluster and P is defined as a core point; Otherwise, P is labeled as a noise.

iii For the points within the ϵ-neighborhood of P, repeat Step i until all the points in the cluster are founded.

iv Choose another unvisited point and repeat all the steps until every point in S are defined.

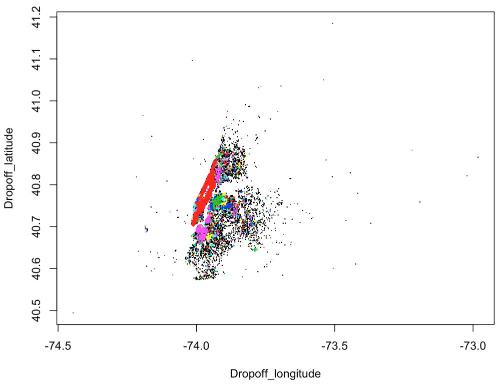

Figure 3.2: DBSCAN Clusters

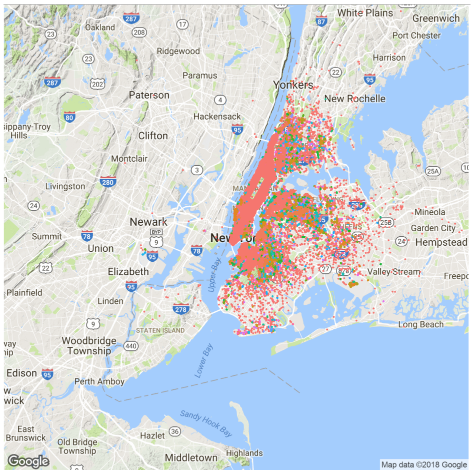

Figure 3.3: DBSCAN Clusters on Map

Figure 3.2 and 3.2 display the results of DBSCAN clustering with $\epsilon = 0.001$ and $minPts = 3$, which is the common choice of $dim + 1$. From the graphs, it is easy to tell that some of the clusters are well defined while some are really small. Also, there exist a large number of outliers.

These short-backs can result from the disadvantages of DBSCAN. First, DBSCAN cannot well cluster data sets varying in densities, since the $minPts - \epsilon$ combination cannot be chosen appropriately for all clusters. Second, the chosen of distance threshold ϵ is difficult when the data and scale are not fully understood. Large $epsilon$ will result in clusters that are too big. If ϵ is too small then there will be too many noise points.

3.2.2 OPTICS

Various extensions to the DBSCAN algorithm have been proposed, including methods for parallelization, parameter estimation, and support for uncertain data. The basic idea has been extended to hierarchical clustering by the OPTICS algorithm. It addresses the weakness of DBSCAN, i.e. the difficulty of identifying clusters varying in densities. In the OPTICS algorithm, the points in the data set are ordered such that those spatially closest become neighbors in the ordering. Besides the core distance (same as in DBSCAN algorithm, the distance from the core point P to the $minPts$th closest point within its ϵ-neighborhood) another distance called reachability distance is computed for each point Q in the ϵ-neighborhood by $max\{core - distance, dist(P,Q)\}$. It represents the density that needs to be accepted for a cluster in order to have both points belong to the same cluster. This is represented as the reachability plot (a special kind of dendrogram). Since the points in the same cluster have a low reachability distance to their nearest neighbors, the clusters are represented as valleys in the reachability plot. The deeper the valley, the denser the cluster (14).

There are multiple ways of extracting clusters from the reachability plot. First, we can manually select a range on the x-axis by observing the intervals. Second, we can choose a threshold on the y-axis. Third, we can use other algorithms to detect the valleys by steepness, knee detection, or local maximums. This method will require multiple runs since the clusters obstained are usually hierarchical.

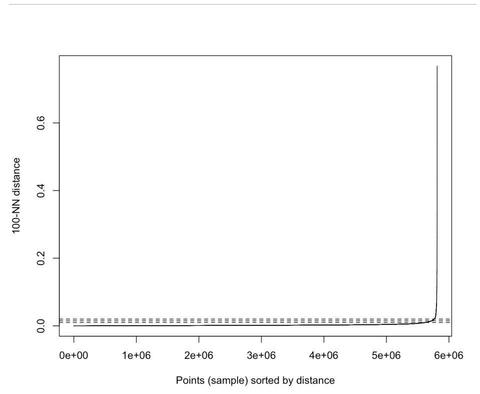

Figure 3.4: $\epsilon = 0.01, 0.015, 0.02$ with $minPts = 100$

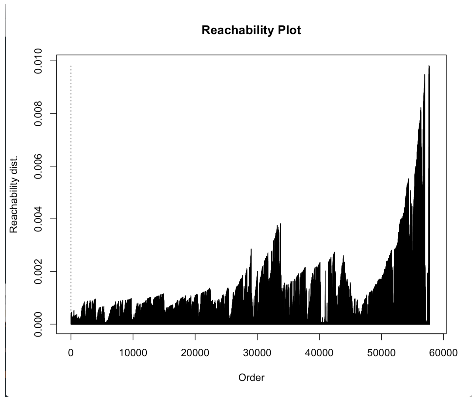

Figure 3.5: Reachability Plot from OPTICS clustering with $\epsilon = 0.02$ and $minPts = 100$

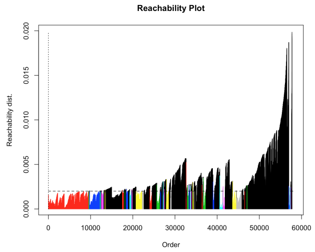

Figure 3.6: Choice of threshold = 0.002

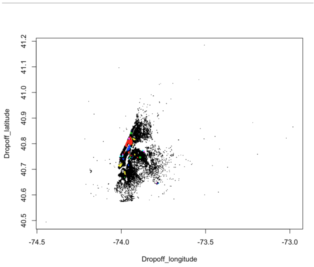

Figure 3.7: OPTICS Clusters with ϵ-threshold = 0.002

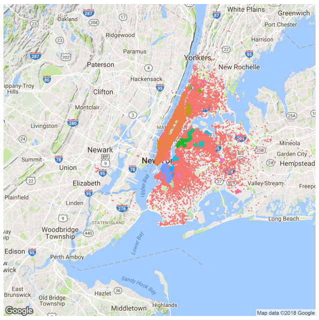

Figure 3.8: OPTICS Clusters with ϵ-threshold = 0.002 on Map

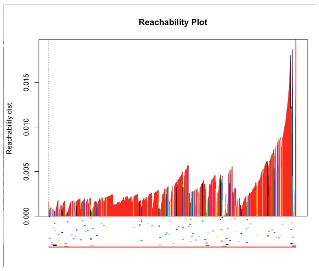

Figure 3.9: Choice of steepness = 0.01

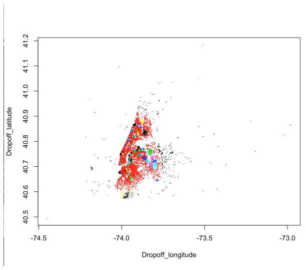

Figure 3.10: OPTICS Clusters with steepness = 0.01

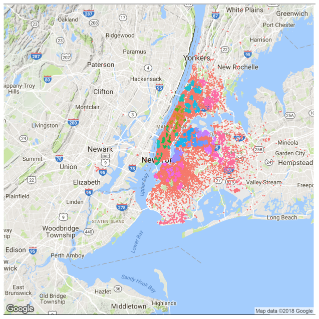

Figure 3.11: OPTICS Clusters with steepness = 0.01 on Map

3.3 Discrete Poisson Point Process

Poisson process is one of the most important and well-studied random processes in probability theory. In probability, statistics and related fields, a Poisson point process or Poisson process (also called a Poisson random measure, Poisson random point field or Poisson point field) is a type of random mathematical object that consists of points randomly located on a mathematical space. The point process has some convenient mathematical properties, which has contributed to its frequent definition in Euclidean space and used as a mathematical model for seemingly random processes in a number of disciplines. For example, it is widely used to model random points in time and space, such as the times of radioactive emissions (15), the arrival times of customers at a service center (16), and the positions of flaws in a piece of material (17). For a Poisson process defined on some underlined space, the number of points N in a bounded subset of the space will be a Poisson random variable with parameter Λ and the probability N being equal to n is:

$$\mathbb{P}(N = n) = \frac{\Lambda^n}{n!} e^{-\Lambda} \tag{3.1}$$

The parameter, usually referred as rate or intensity, is related to the expected (or average) number of Poisson points existing in some bounded region, and rate is usually used when the underlying space has one dimension (18). The parameter Λ can be interpreted as the average number of points per some unit of extent such as length, area, volume, or time, depending on the underlying mathematical space, and it is also called the mean density or mean rate (19).

In order to establish a quantified relationship between the pick-up and drop-off locations taking account of the time, I generated a discrete Poisson process model. Different from the regular continuous Poisson process defined on the real line, where at each interval t_i, only one event (here one trip) occurs, I implemented the model in a discrete setting, in which multiple trips, N_t happen at different pickup time t. Therefore, in my model , λ denotes the mean density of trips across the time in a day. Besides, I have assumed that the pickup locations occurred on each unit point in region S is equal which results in a Uniform distribution in the bounded region S. Since the pickup times have been set to every second in a whole day, or $T = 24 \times 60 \times 60\ seconds$, we can say it follows a Uniform distribution on the $[0, T]$ interval. Hence, we could have the following densities:

$$P(N_t) = \frac{e^{-\lambda} \lambda^{N_t}}{N_t!} \tag{3.2}$$

$$P(\theta_{ip}) = (\frac{1}{Area(S)})^{N_t} \tag{3.3}$$

$$P(t) = (\frac{1}{t})^{N_t} \tag{3.4}$$

Noted that, in this part, I used the original time data instead of the parsed hourly data and the original (longitude, latitude) coordinates without rounding them.

3.3.1 Model Learning

Suppose the number of trips N on the time interval $[0, T]$ on the 2-D space, the area of New York City S. Generate the number of trips N_t at time t by Poisson distribution with parameter λ, which can be interpreted as the average number of trips per unit time interval:

$$N_t \sim \mathcal{P}oi(\lambda) \tag{3.5}$$

Consider each N_t trips. Generate the pickup locations $\theta_{1p}, \theta_{2p}, \ldots \theta_{N_tp}$ and pickup times $t_{1p}, t_{2p}, \ldots t_{N_tp}$ by Uniform distribution on the area S and total time T:

$$\theta_{1p}, \theta_{2p}, \ldots \theta_{N_tp} | N_t \sim \mathcal{U}(S) \tag{3.6}$$

$$t_{1p}, t_{2p}, \ldots t_{N_tp} | k \sim \mathcal{U}(0, T) \tag{3.7}$$

From Figure 1.1, we have already learned that the trip duration follows a Gamma distribution with parameter μ, which can be interpreted as the inverse of average trip duration. Therefore, we can generate the drop-off times $t_{1d}, t_{2d}, \ldots t_{N_td}$ accordingly. Simultaneously, we can generate the drop-off locations $\theta_{1d}, \theta_{2d}, \ldots \theta_{N_td}$ by a joint Bivariate Normal distribution with parameter Σ representing the radius of trips from a given pickup location:

$$\theta_{id} | \theta_{ip} \sim \mathcal{N}(\theta_{is}, \Sigma_i) \tag{3.8}$$

$$t_{id} | t_{ip} \sim Exp(\mu) + t_{ip} \tag{3.9}$$

Deriving from 3.4, we can further extend the distribution of pickup and drop-off locations into a distribution of the trip distance (here I used Euclidean distance to estimate σ^2):

$$\mathbb{E}||\theta_{id} - \theta_{ip}||^2 \sim \sigma^2 \mathcal{X}^2(2) \tag{3.10}$$

In this case, we can use σ^2 as a simplified measure of the covariance between pickup and drop-off locations.

3.3.2 Model Fitting

From the original data, we can estimate λ using the average number of trips per unit time (per second). Since I am using a discrete Poisson process, this average rate will be equal to the number of trips starting at each second. The parameter μ for the exponential distribution of trip duration can be estimated from the inverse of the average trip duration throughout the day (across all the trips). In the joint normal distribution where we fit the pickup and drop-off locations, we can evaluate Σ by the covariance of $\theta_{id} - \theta_{ip}$. From 3.6, variance σ^2 can be evaluated from the average of the square of trip distance:

$$\hat{\lambda} = \frac{Total\ number\ of\ trips}{24 \times 60 \times 60} \tag{3.11}$$

$$\hat{\mu} = \frac{1}{avg\ (Trip\ duration)} \tag{3.12}$$

$$\hat{\sigma^2} = \frac{\Sigma\ dist^2}{2 \times Total\ number\ of\ trips} \tag{3.13}$$

Plugging the data into 3.11, 3.12 and 3.13:

$$\hat{\lambda} = 17.043 \tag{3.14}$$

$$\hat{\mu} = 0.0008182 \tag{3.15}$$

$$\hat{\sigma}^2 = 0.001210555 \tag{3.16}$$

Next, I simulated new data from the above distribution and estimators, from which I achieved new values of the parameters λ, μ and Σ. In order to evaluate the fittness of the model we just generated from the above distributions in 3.1 - 3.6, I compared these new values of the parameters with the estimators.

3.3.3 Model Evaluation

From the simulated data, we get $\lambda = 17.0384$. For the parameters μ and σ^2, we get a value for each n_i trips starting at each second t_i. To compare these values with the estimators, I plotted the Figure 3.12 and 3.13, in which the red lines represent the value of estimators while the black dots are the parameter values we computed from our simulation:

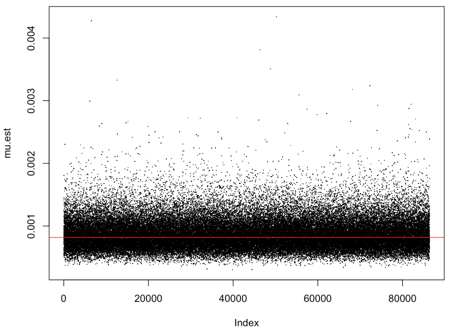

Figure 3.12: μ from Simulated Data

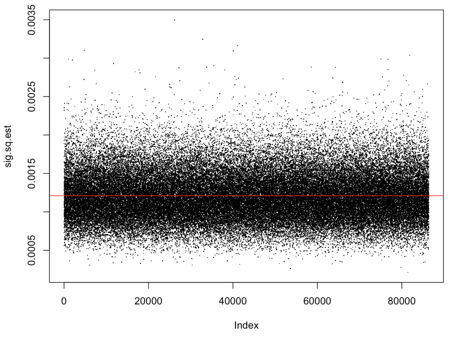

Figure 3.13: σ^2 from Simulated Data

From the above figures, we can see the the allocation of values of μ and σ^2 from the simulated trips.

In order to compare the locations of the pickups and drop-offs, I plotted the simulation on the map. Since I generated the pickup locations from a uniform distribution on a certain area of New York City (a rectangle formed by constrains of longitudes and latitudes), the points of pickup locations formed a rectangle on the map and made it hard to see the locations of drop-offs. Therefore, I randomly choose 1000, 5000, 10000 trips from the simulation and compared them with the original location distributions.

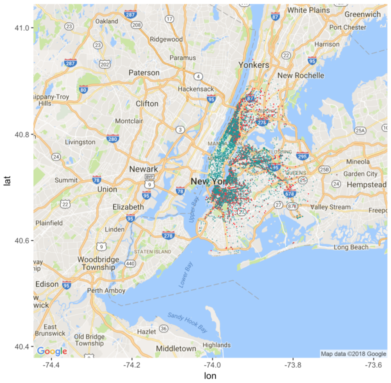

Figure 3.14: Random 5000 trips from the original data

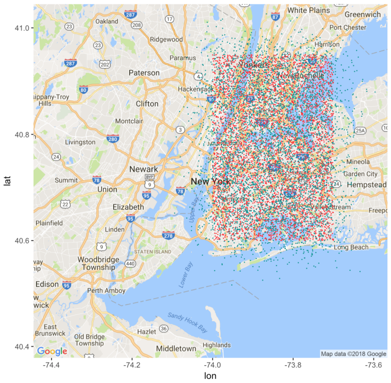

Figure 3.15: Random 5000 trips from the simulated data

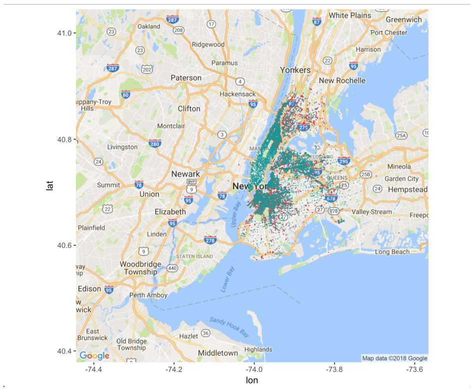

Figure 3.16: Random 10000 trips from the original data

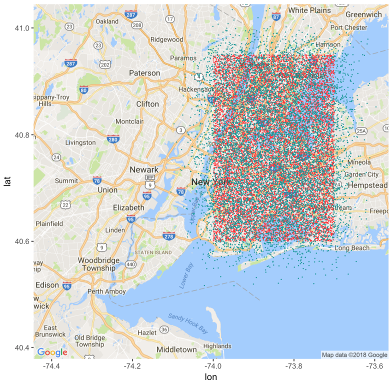

Figure 3.17: Random 10000 trips from the simulated data

CHAPTER 4

Conclusion

4.1 Summary

Through the analysis of the NYC TLC taxi data, the overall condition of the traffic flow on taxi trips has become clearer. The common directions from a certain pickup region to a drop-off region can be detected and used to show the most popular routes of taxi passengers at different time of the day. Through the analysis of the data with spatial taxi zones, we have learned that a lot of trips went on within a certain zone. Hence, we could conclude that these zones are the ”busiest” regions where people travel with a taxi. The change of these zones across time in a day showed us the variation in starting points and destinations throughout the day. It gave us a preview of potential function zones in the New York City. For example, during the morning, it tends to be the office or school districts that most people head to while later in the afternoon it becomes the living neighborhood. Further more detailed analysis can be done with supporting information of the real estate of different zones, from which we can possibly learn the relationship between the economic background and passengers' travel patterns.

The division of traffic based on the density of trips could be helpful to understanding the commuting patterns, and thus, can serve as a reference for urban planning and transport optimization problems. It could also be valuable as a method of analyzing traffic congestion and to reorganize or redesign the urban traffic system for a more efficient and less consuming traffic network. Even though manually bounding traffic circle is a time and labor consuming procedure, it provided a deeper insight of how the traffic was like from the data being provided. The density based clustering did a better job in detecting clusters, which represents traffic dense circles, in irregular shapes other than round circles. It was faster and more precise in find the centers of those traffic circles. However, due to the fact that there were a lot of variations and outliers, it was hard to include all the points into the clusters. Nevertheless, its assistance in detecting the cluster centers was still very helpful.

The bounding technique could serve as a supplement to the density based algorithms. Some possible solutions can be: first use the density based clustering algorithms to find out the center of the traffic circles; then use those centers as the centroids of the bounding circles.

4.2 Limitation

From Section 2 we have learned that there are a lot of common routes locating within a taxi zone. However, there is no information or measurement we can use to further analyze these inner-zone-trips. However, if we can get the range of longitudes and latitudes of each taxi zone, we can used the data from previous years to conduct further analyses. Therefore, there needs to be a method that can transform between exact geological coordinates and spatial zones.

From Figure 3.15 and 3.17 we can see that since the pickup locations are generated from a Uniform distribution on a bounded rectangle box on the geological 2D plane of longitudes and latitudes, it is hard to find the most suitable boundaries for this box, and thus, it resulted in the situation that many of the location points were located to regions where the taxi trips could not happen at all.

4.3 Future Steps

My finished work on the problem is relatively simple and has many limitations as stated above.

i Further extend the density-based clustering and find efficient way to extract clusters.

ii Implement the Poisson Point Process. Let λ be a function on t instead of a fixed constant.

iii Divide the map of New York City into small partition boxes and generate pickup locations on each small partition to compute the parameters. Then compute the σ^2 for each small partition and compare them. The smaller the σ^2, the busier the partition. Then we can define traffic circle based on the partitions with small σ^2's.

BIBLIOGRAPHY

[1] HOW MANY CARS PER CAPITA IN THE USA.

[2] Walden, C. (2015) Fundamentals of Transportation, CHIZINE PUBN, .

[3] Cox, W. (2010) NEW TRAFFIC SCORECARD REINFORCES DENSITY-TRAFFIC CONGESTION NEXUS.

[4] Bernard, M. TRAFFIC CONGESTION: HOW PREDICTABLE? Discovering Volume Trends across Time and Confirming Fundamental Speed-Flow-Density Relations. (2015).

[5] Lu, H.-P., S. Z.-Y. . Q. W.-C. (2015) Big Data-Driven Based Real-Time Traffic Flow State Identification and Prediction. *Discrete Dynamics in Nature and Society,* **2015**(284906), 11 pages.

[6] INRIX Global Traffic Scorecard. (2017).

[7] Deville, P., L. C. M.-S. G. M. S.-F. R. G. A. E. e. a. (2014) Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences,* **111**(45), pp. 15888–15893.

[8] Khodabandelou, G., G. V. E.-Y. M. . F.-M. (2016) Population estimation from mobile network traffic metadata.

[9] TLC Trip Record Data. (2017).

[10] Turner, S. (1998) Travel time data collection handbook, Washington, DC: Office of Highway Information Management, Federal Highway Administration, U.S. Dept. of Transportation, .

[11] Al-Sobky, A.-S. A. and Mousa, R. M. (2016) Traffic density determination and its applications using smartphone. *Alexandria Engineering Journal,* **55**(1), 513 – 523.

[12] May, A. (1990) Traffic flow fundamentals, Englewood Cliffs, N.J. : Prentice Hall, .

[13] Roess, R.P., P. E. . M. W. (2004) Traffic Engineering, Pearson Prentice Hall, 4 edition.

[14] Ester, M., K. H. S. J. and Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining,* **96**(34), pp. 226–231.

[15] Buczyk, B. (2009) Poisson distribution of radioactive decay.

[16] Shen, H. Huang, J. (2008) Forecasting time series of inhomogeneous Poisson processes with application to call center workforce management. *The Annals of Applied Statistics,* **2**(2), pp. 601–623.

[17] Composite materials : testing and design (fourth conference) : a conference Philadelphia, Pa. : ASTM (1977).

[18] Kingman, J. (1992) Poisson Process, Clarendon Press, .

[19] Daley, D.J. Vere-Jones, D. (2007) An Introduction to the Theory of Point Processes: Volume II: General Theory and Structure, Springer Science Business Media, .