Anomaly Detection in NY Taxi Data Using LRT

Harish Pullagurla
Department of ECE
North Carolina State University
Email: hpullag@ncsu.edu

Hari Krishna Majety
Department of CSC
North Carolina State University
Email: hmajety@ncsu.edu

Kenneth Tran
Department of CSC
North Carolina State University
Email: kvtran2@ncsu.edu

Abstract—**In this work, we propose modifying the Likelihood Ratio Test (LRT) for anomaly detection by re-defining how to sub-sample the data. By sub-sampling in an intelligent way, we reduce the number of possible anomalous segments we test for. We perform a spatial and a temporal analysis separately to determine both anomalous regions and anomalous times within the NYC taxi cab data set provided by Kaggle. In our spatial analysis, we conclude that anomalous regions, such as the JFK airport, are properly detected by the framework. In our temporal analysis, we detect days of the year in lower Manhattan that are anomalous. We cross-check our findings with special events in lower Manhattan to ensure the anomalies being detected are significant.**

I. Introduction

In 2015, ECML-PKDD had a competition for predicting taxi trip times. The purpose of this was to improve the efficiency of electronic taxi dispatching systems. The idea is that if a taxi dispatcher knows approximately when a taxi will finish its trip, then it can immediately assign it to another nearby trip. Besides efficiency within the companies, taxi data can also be used to observe the current state of traffic in a given region. Since taxis use the same roadways as other vehicles, their trip duration are indicative of congestion within regions. This type of analysis is related to urban computing, where computation is used for urban planning and development.

One of the biggest goals in urban computing is to decrease traffic congestion. Locating where traffic is heaviest or lightest are important in alleviating congestion. For example, if a GPS software knows where traffic is light, it can direct drivers moving towards heavy traffic to areas that are not congested. Many machine learning algorithms can be applied to automatically capture these patterns in traffic data. In particular, various methods of anomaly detection have been created to find the source of congestion to help policy makers determine how to deal with the issue. In our project, we modify the anomaly detection likelihood ratio framework by Yang et al. [2] to find anomalies in the NYC taxi cab data. In particular, we propose modeling the trip duration provided by the Kaggle NYC taxi cab data set instead of purely using trip counts. We also propose reducing the number of regions to compute by performing clustering during pre-processing.

II. Related Works

In [1], Wu et al. uses the likelihood ratio test (LRT) to detect anomalies. For the LRT, two models are compared by taking the ratio of their likelihoods, which is computed using a function given by the user. Parameters for the two models are computed using maximum likelihood estimation. To extend this to anomaly detection, they compare the likelihoods of two different models. The first model is one that fits a set of parameters over all of the data, which acts as our null hypothesis. The alternative hypothesis is that the data is better fitted using two separate models, one fitted on the anomalous subset and one fitted on the remaining data. Just like for the LRT, we take a ratio of these two options, combining the anomalous data model and remaining data model by taking a product of their likelihoods. The problem with the classical LRT for all contiguous regions given an nxn grid, there are n^4 possible regions to check, which makes the computation cost extremely high. To deal with this problem, the authors propose a pruning algorithm that first filters out all candidate regions that are not anomalous. The intuition behind this is very similar to the A Priori algorithm for frequent item set mining. They make the observation that the likelihood of a given region R (which consists of R_1 and R_2) is always less than the product of the likelihood of R_1 and the likelihood of R_2, where R_1 and R_2 are disjoint. Therefore, if the likelihoods of a regions sub-regions are pre-computed, then their product can be compared to a threshold value.

Yang et al. [2] applies this framework directly to Beijing taxi cab data. Their goal is to identify a set of contiguous blocks inside the region which were anomalous. In this case, they used a grid structure to split up Beijing into regions to set up for using the pruning mechanism developed in the framework. In taxi cab data, temporal attributes play a significant role. Therefore, they proposed a method to split up regions in the grid on the time axis as well, resulting in cubic regions. For this set-up, the computation complexity given an nxn spatial grid with t time segments would be n^4t^2 using brute force. By using pruning mechanisms, they reduce this to n^4. They also introduce the temporal aspect into the spatial model by proposing an optimization of parameters with respect to the time attribute in the data. They used two case studies, one corresponding to the Chinese holiday "Golden Week" and one during the National People's Congress, both in 2009. Their results found that for "Golden Week", the popular attractions for this holiday resulted in anomalous activity and that for the National People's Congress, a road left open during the meeting became anomalous, probably because other common roads were blocked.

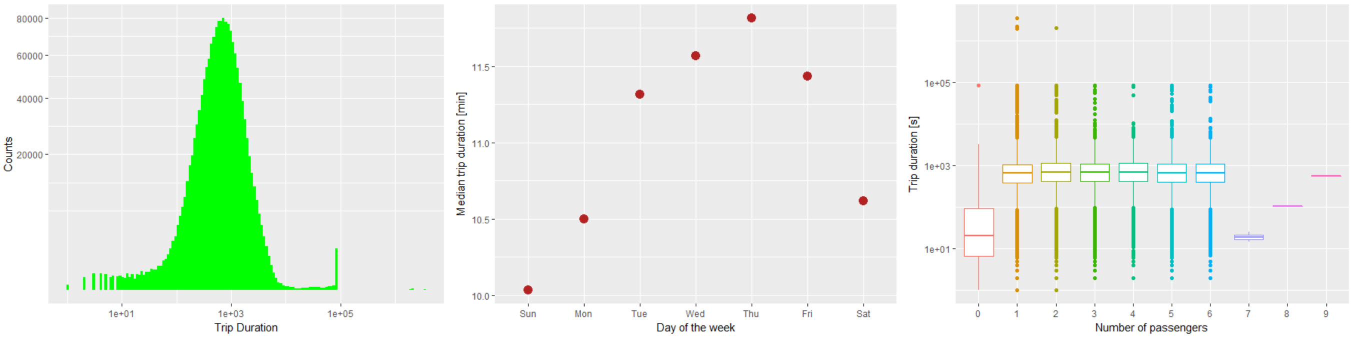

Fig. 1. [A] Logarithmic plot of time taken vs number of trips, Variation in Trip Duration with [B] day of Week, [C] passenger count

III. Methodology

A. Data Sets

Yellows cabs play a very important role in customer's daily commute in New York City area. The NYC Taxi Limousine Commission[3] has publicly made available the Trip Record Data (TRD) for all trips made in the city starting 2009. Kaggle, a platform for data analysis, has hosted the NYC trip duration estimation challenge[4], built around the same data set. For the hosted competition they provide cleaned and sampled data from first half of year 2016 with an expectation to predict the travel time. The data set has around 1.5 million trip records, each with 11 feature attributes. For experimentation in our work, the data hosted in this competition has been used.

The base features give information about different aspects of a trip :

1) Spatial Information [4 attributes] Pick up and drop off - Latitude and Longitude information
2) Temporal Information [3 attributes] Pick up, drop off date and time along with trip duration
3) Other General Information [4 attributes] unique trip id's, vendor id, passenger counts, Information of transmitting it to the server 4 attributes

B. Exploratory Data Analysis

Exploratory analysis on the original data-set is done to understand the patterns in the data in more detail.

Figure 1, [A] shows that majority of rides follow a normal distribution getting a peak at around 10000 sec. It can be seen that there are some very short duration trips which 10 sec duration. Also a surprising peak is observed at the end of the distribution, which could be the trip from/to airports from Manhattan.

It can be observed from figure 1[B], the variations of trip duration with day of week. Trip Duration appears to be high and almost constant during the working weekday except for Monday. Also, as expected the trip duration is smaller for weekends.

Passenger count gives the number of people traveling in the cab during the trip. Figure 1[C] shows that, trip duration for passenger counts 1 to 6 are almost similar. Points with passengers 0 is strange, and it could probably mean the taxi travelled without any passengers.

From the plot in figure 2, it could be seen that trip duration is fairly high and constant during the working hours 8 am onwards . The timings seem to have a downward trend during the later parts of the day post midnight. It is in lines of the general notion that most people travel during the day mostly during the work hours and generally there is a lower amount of traffic.

C. Data Prepossessing

The data set used, was clean to a large extent without any major missing attributes. Full data point was dropped if there were any missing attributes in the data set. Uniqueness of all id's was checked and pruned accordingly. Few data points were observed to be having pick up and drop locations in the sea and places as far as California. Latitude and Longitude information of interest, in this experiment, is in and around the city limits of New York. Any other points will be considered erroneous, as the measurements from those trip would either have large travel times or were from faulty meters, hence not fitting the general trend of data. A bound on latitude and longitude with limits in the range of [40.63 , 40.85] and [-74.25, -73.77] respectively were set. Doing this step reduced the number on data points to 1.438 million samples from 1.45.

Few trip durations and passenger counts that appeared seemingly strange more like some erroneous entries rather than anomalies. We assume any point which is greater than 5 standard deviations away from the sample mean to be an erroneous entry and removed them from experimentation. Passenger count of 0,7,8 9 was seen to be highly different compared to the regular valid plots and also is seen to be present in only a small number of data points. Hence these points have also been ignored for experiments in this project.

D. Feature Addition

Several new features were added to the original data set which has 11 attributes on trip record information, based on the modelling demands encountered. New Features considered are as follows:

OSRM Data: To improve the real world travel nature of trips, data queried from Open source routing machine (OSRM) [5] was considered. OSRM is similar to google maps. From this query data, 3 attributes, 'total trip distance', which provides on road distance between the pickup and drop locations,

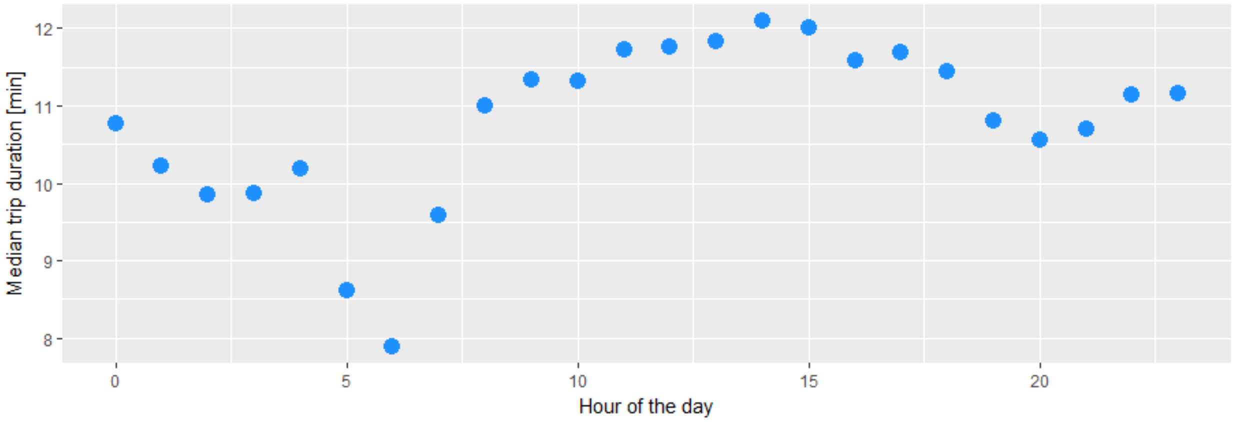

Fig. 2. Variation in Trip Duration with hour of the day

'total travel time', which tells about the general estimate of travel time and 'number of steps', giving information about turns to be taken during the trip, were added.

Date Time Information: Date and time of pick up and drop were provided as strings. Several features were extracted from the this string for easy interpretation of timings. Day of year, day of week, hour of day, week of year were few of the attributes extracted from the date time string for easy sub sampling of data while experimentation.

Haversine Distance: The haversine formula determines the great-circle distance between two points on a sphere given their longitudes and latitudes. It gives the aerial distance for the data points.

Region Clustering: The city needs to be sub divided into smaller regions for further region based analysis. A k means based clustering technique with 20 clusters was used for this process. Clustering was done on pick up locations of these points and drop location used the same parameters. Figure 3 [A,B] shows the initial plot and label outputs obtained for the pick up locations.

It was observed that clusters not in Manhattan, or near the Airports were sparsely distributed. The number of points in these clusters were less than 10% of the mean number of data points in other regions. These clusters were ignored in this experiment, as the spatial modeling was based on the assumption that all clusters are similar. Clusters with these locations were ignored and the remaining points were passed through a new k means prediction method to get fresh cluster labels. Label maps for pick up and drop off locations are added as 2 new attributes to the data set. Figure 3 [C] shows the label map for different locations, which will be used for all future analysis.

E. Sub Sampling

Different modeling strategies were considered for this project which work on certain under-laying assumptions. The models in common assume that most of the data is similar in different subsets of it. If they are any differences they would account for the anomalous measurements. An attempt was made to reduce the data-set of any such known extreme variations which would question the validity of the models . The data set was sampled to include only weekdays from Tuesday to Friday. It is seen from the exploratory plots in Figure 1 that these fours days follow a similar trend in comparison to the other data points.

On similar lines, data was also sampled by hour of day. Figure 2 shows that the variation over trip duration, remains constant during 8am to Mid night. Other sample points are ignored from the experimentation for making the data free from the known trend.

Also for the temporal models a sub sample Lower Manhattan Region from the full region was considered. This was done to only include the regions which had similar density/pattern, which was based on outputs obtained from spatial model for anomaly detection.

F. LRT Framework

In our work, we propose simplifying the computation of the LRT based on large data sets by using intelligent partitioning methods instead of using pruning, as proposed by Wu et al. [1] In our first model, we take into account the spatial differences and look at anomalous regions in the data. In our second model, we use temporal differences and instead for look anomalous days in a given region. For both models, the data is split up into either regions or times. We will use the general term *segments* to encompass both.

In our models, we employ the anomaly detection LRT framework. Here, we will explain the method in more detail. In this test, the null hypothesis assumes that there is no difference between the current segment compared to the rest of the segments in the data. This means that all the data is similar and can be modeled using one set of parameters. The alternative hypothesis is that the current segment is anomalous compared to the rest of the data. This means that the subset of data for the current segment should be modeled with a different set of parameters from the rest of the data. For the LRT, we take the

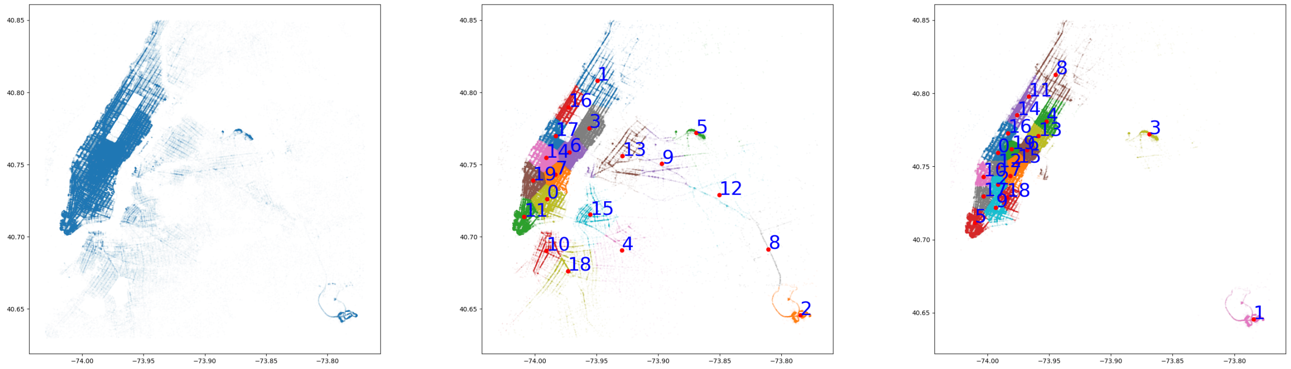

Fig. 3. x,y axis = latitude and longitude [A] Pickup Positions, [B] k means Cluster labels, [C] Cluster labels post cleaning

ratio of the product of the two likelihoods from the alternative hypothesis to the likelihood of null hypothesis.

Mathematically, the likelihood ratio can be described as:

$$\lambda = \frac{L(\theta_s|X_s)L(\theta_{\bar{s}}|X_{\bar{s}})}{L(\theta|X)} \tag{1}$$

To obtain the model, we use generalized linear models (GLMs) using maximum likelihood estimation (MLE) to obtain the optimal parameters. We used the trip duration as our predicted variable and OSRM distance, OSRM time, and Haversine distance as our predictor variables. Given that trip duration is continuous, we chose to test our models using the Gaussian family with an identity link (normal linear regression) and the Gamma family with a log link (a fairly common GLM configuration). Then, based on the parameters from MLE, we obtain the overall likelihood of each model given the data. Due to the extremely large differences in sample size between the global model and the segment models, we use the average log-likelihood.

Once the models are fitted, we can get the test statistic using the following:

$$\Lambda = -2log(\lambda) \tag{2}$$

By comparing test statistic between different segments in the data, we can determine which ones are anomalous in comparison to others. We also show a visualization of what the LRT is attempting to do in the case of linear regression in Figure:4. For this example, the graph on the left attempts to fit a line on the data globally. The graph on the right then fits two lines separately on the orange group of data and the blue group of data. Due to the global data being poorly fitted on the left graph, the likelihood ends up being low, while the likelihood of the two models on the left are high. Therefore, the ratio between the product of the two high likelihoods and low likelihood (global model) will be high, indicating that the sub-sample is an anomaly.

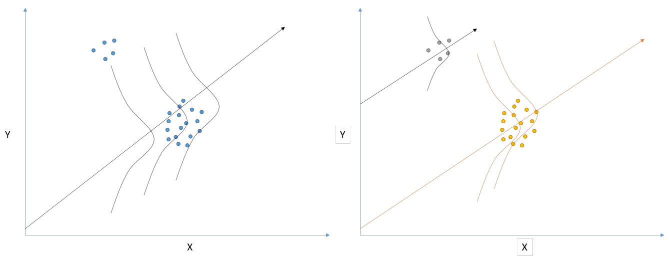

Fig. 4. [A] LRT Null (No Anomaly), [B] LRT Alternative (Anomaly)

1) Model 1 - Spatial Analysis: For the first model, we perform spatial analysis on our data. Our aim here is to determine which regions of NYC are more anomalous than others. This means that our segments will be defined as regions in the data. To obtain the regions, we took the clusters created by the K-Means clustering we performed in pre-processing. For the alternative hypothesis, we build a model on only data with pick up locations in the current region and a model on only data with pick up locations outside of the current region. While researching the data for this project, we found that airports were likely to be anomalies in the data. We implemented this more basic model as verification for our framework.

2) Model 2 - Temporal Analysis: For the second model, we perform temporal analysis on our data. Our aim here is to detect which days were more anomalous. If we have this information, we can infer that something happened (an event, large accident, etc.). This means that our segments will be defined as time slices in the data. TO obtain the time slices, we used the variable day of year. Since our training data consists of half of the year 2016, this means that there are about half a year of days, 183. Our intuition was that if we did this analysis over all of NYC, our results wouldn't be as meaningful. Instead, we sub-sampled the data to only look at lower Manhattan. Looking at only one particular region decreases the variability in our data, which effectively makes anomalies more detectable and significant.

IV. Implementation

Code was primarily written in python 2.7 using numpy, scipy, statsmodel packages. Visualizations were plotted in R.

The spatial and temporal analyses were performed using Generalized Linear Model of the statsmodels.api package in python and the analyses were run with both Gaussian and Gamma distributions. The likelihood resulting by plugging in the regions and fitted parameters is then considered for the Likelihood ratio test and we pick the top 5 regions or times depending on the ratios which are relatively anomalous than the other regions.

```
**Data:** pre-processed filtered data ($X$), segments in data ($S$), GLM family and link
**Result:** Likelihood ratio value vector (allLLR) for all $s$ in $S$
Obtain null model using MLE (get $\theta_0$)
Compute log-likelihood of null model ($L(\theta_0|X)/|X|$)
**for** $s$ *in* $S$ **do**
    Obtain current alternative model using MLE (get $\theta_s$ and $\theta_{\bar{s}}$)
    Compute log-likelihood of current alternative model ($L(\theta_s|X_s)/|X_s|$ and $L(\theta_{\bar{s}}|X_{\bar{s}})/|X_{\bar{s}}|$)
    Compute likelihood ratio ($\lambda_s = \frac{L(\theta_s|X_s)L(\theta_{\bar{s}}|X_{\bar{s}})|X|}{L(\theta|X)|X_s||X_{\bar{s}}|}$)
    Append $\lambda_s$ to allLLR
**end**
```

V. Results

A. Model 1 - Spatial Analysis

The spatial analysis with Gaussian distribution on the dataset showed JFK Airport region as the most anomalous region compared to other regions which could be due to frequent commuters to and from the Airport which is relatively far from the city. The next anomalous region happens to be another airport in the city, the La Guardia Airport. These regions were notably different from the rest.

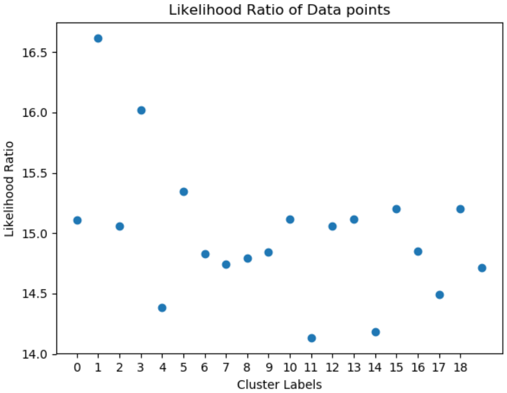

Fig. 5. Likelihood ratios of Spatial Clusters

B. Model 2 - Temporal Analysis

The temporal analysis on the dataset with gaussian distribution showed January 26th as the most anomalous day in the year which happened to be the time when New York was hit by biggest blizzard since 1869[8]. This was followed by June 8 being next most anomalous day in the subset of data we chose. June 8 happens to be the day when President Obama taped a show with Jimmy Fallon in NYC[6]. The next anomalous most anomalous day was May 16, the day when there was a Fire under Metro-North tracks which was caused by a fuel spill [7]. The same analysis when run with Gamma distribution also gave the same output.

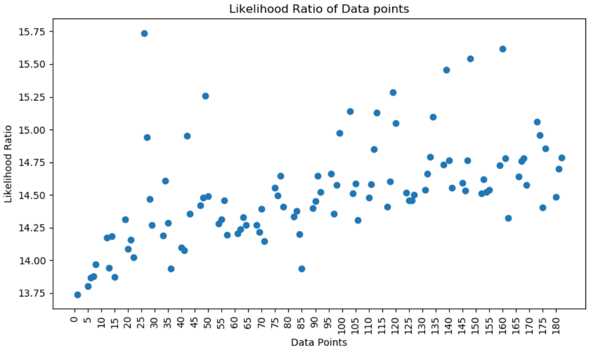

Fig. 6. Likelihood ratios of Temporal Clusters

VI. Conclusion

In this work, we develop a likely-hood ratio test based anomalies detection algorithm. We test the algorithm separately in spatial domain and temporal domain. The detected anomalies were cross verified with actual events and were seen to have a meaningful correlation. A possible extension of this work could be a spatio-temporal analysis where we take both spatial and temporal aspects into account while computing the likelihood of a particular data point. This extension makes the computations complex requiring computational resources.

At present we work with an assumption that all data subsets are similar, which is not true if data from weekends and night hours are taken into account. As a next extension seasonality of the data-set can be taken into account while building the model. Also, the model is built upon likelihoods from data samples. When sub-sample sizes are different, they cannot be directly compared, hence averaging was done. But from 6 it is seen that there is a correlation between number of data points and its values, which is to be addressed in future work.

Repository

Link to Implementation is present on a Github Repository: https://github.com/pharish93/NYTaxiDataMining

References

[1] Wu, Mingxi, et al. A LRT Framework for Fast Spatial Anomaly Detection. Research Gate, Proceedings of the 15th ACM SIGKDD, Jan. 2009.

[2] Pang, Linsey Xiaolin, et al. On Detection of Emerging Anomalous Traffic Patterns Using GPS Data. Data Knowledge Engineering, North-Holland, 18 May 2013.

[3] Trip Record Information, publicly available by the NYC Taxi Limousine Commission. URL: http://www.nyc.gov/html/tlc/html/about/triprecorddata.shtml , (last accessed: 04.11.2018)

[4] Kaggle NYC taxi trip duration estimation Challenge, URL: https://www.kaggle.com/c/nyc-taxi-trip-duration , (last accessed:04.11.2018)

[5] Open Source Routing Machine, URL: http://project-osrm.org/, (last accessed:04.11.2018)

[6] Obama guidance, Press Schedule June 8,2016 , URL: https://chicago.suntimes.com/news/obama-guidance-press-schedule-june-8-2016-to-nyc-jimmy-fallon/, (last accessed:04.18.2018)

[7] Fire Under Metro-North tracks, URL: https://www.nytimes.com/2016/05/19/nyregion/metro-north-service-fire.html, (last accessed:04.18.2018)

[8] NYC Blizzard breaks records, URL: https://patch.com/new-york/williamsburg/nycs-first-blizzard-2016-breaks-snowfall-record, (last accessed:04.18.2018)

ACKNOWLEDGMENTS

We would like to thank the course administrators of Spatial and Temporal Data Mining, Spring 2018, NCSU, for providing us the platform to work with a really good team and an interesting problem statement.

PROJECT CONTRIBUTIONS

Common : Time was spent individually, during the initial phase, in reading the existing work in the field of anomalies detection and to understand work related to NYC Taxi data analysis. In the second phase, time was spent together in meetings (around 30 - 35 hours), which were mainly intended to formulate the theory for this specific problem statement. Here, we explained each other the learning's from this initial phase and how it needs to be modified for this data set. This was interleaved with code implementation and experimentation done individually. In the final phase, time was spend in documenting the observed results.

Harish Pullagurla : During the initial phase, I spent time to understand the data set and different patters in it. Information available on several kaggle kernels was a starting point in this process. This knowledge was useful while formulating the problem and in helping us take a step by step approach towards developing the present solution. I handled setting up the basic framework for the code along with writing the data prepossessing, cluster labeling, sub sampling functions and few visualization functions in python.

Hari Krishna Majety: Initially, I spent most of my time in understanding the LRT and other involved statistical and modeling concepts and sharing my understandings with the group. Then I explored the dataset and identified the points which seem to occur possibly due to erroneous entries rather than anomalies and I tried out various parameters and methods to filter out such points from the dataset. Later I created visualizations of exploratory analysis plots in R.

Kenneth Tran: I spent a majority of my individual time understanding and explaining the papers on the likelihood ratio test, including the one given in class as a reading and the extension on it to taxi cab data. With helpful discussions with my group (where I created our LRT visualization), I also determined how to best apply the LRT framework to our specific problem. After looking at modeling options, I chose to use the SM python package and employ GLM for our MLE model. Lastly, I wrote the implementation for the LRT and added it to Harish's workflow.