Taxi Usage Modeling with Spatio-Temporal Correlations

Kun Xie, Yuzheng Zhuang, Ya Liu

ABSTRACT

The objectives of this project are: 1) to explore how taxi usage is related with factors such as population, median income, road network and public transit accessibility; and 2) to investigate the spatial and temporal patterns of taxi usage. Census tracts are used as the basic geographic units to capture taxi usage, transportation, land use, demographic and socioeconomic data. A Hadoop MapReduce program is designed to compute the monthly taxi pick-ups and drop-offs for each census tract. Bayesian hierarchical models with spatial and temporal effect terms are used to identify the contributing factors to taxi trips. Seven explanatory variables, which include number of bus stations, number of subway stations, number of bike racks, road density, median income, employment, and population, are found to affect taxi usage significantly. September, May and March are expected to be the top three months that have the highest taxi usage; while the February and November are predicted to among the months that have lowest taxi usage. Additionally, the spatial patterns of taxi trips are investigated and the unobserved heterogeneity among NTA has been confirmed. Understanding the impacts of contributing factors to taxi usage and the spatio-temporal patterns can provide insights to aid in better planning, design and management of taxi operation system.

Table of Contents

List of Figures

List of Tables

INTRODUCTION

The objectives of this project are: 1) to explore how taxi usage is related with factors such as population, median income, road network and public transit accessibility; and 2) to investigate the spatial and temporal patterns of taxi usage. The differentials in socio-economic and transportation features have impacts on the generation and attraction of taxi trips. Uneven distributions of taxi demand and supply across the city is likely to lead to the spatial clustering of taxi pick-ups and drop-offs. Additional, temporal correlation could be presented regarding the change of people's preference to travel. Understanding the impacts of contributing factors to taxi usage and the spatio-temporal patterns can provide insights to aid in better planning, design and management of taxi operation system.

New York City is used as the study area. A massive amount of taxi trip data is generated over time, and it is difficult to be processed and analyzed. MapReduce is a programming model for expressing distributed and parallel computations on large-scale data processing. MapReduce has been widely adopted via an open-source implementation Hadoop. Another focus of this study is to design a MapReduce program to explore the "Big Data" that generated by taxi trips.

PROCESSING BIG DATA

Data Sources

The census tracts of Manhattan were used as the basic geographical units for data preparation and modeling. The geographic information system (GIS) data of census tracts were provided by New York City Department of City Planning (NYCDCP[1]). The census tracts could be easily connected to the census data provided by U.S. Census Bureau[2]. Demo-economic features including population, employment and median income of census tracts were collected.

Four-year New York City taxi data from 2010 to 2013 was obtained from the website of Prof. Dan Work[3] who requested it from New York City Taxi & Limousine Commission (NYCTL). This dataset includes 697,622,444 trips of yellow taxi cabs in New York City. The monthly pick-

[1] Source: http://www.nyc.gov/html/dcp/html/bytes/districts_download_metadata.shtml
[2] Source: http://factfinder.census.gov
[3] Source: http://publish.illinois.edu/dbwork/open-data

ups and drop-offs were calculated for each census tract. A MapReduce program was designed to process the massive taxi dataset, more details of which are presented in the next section.

The GIS data of bus and subway stations were obtained from the Metropolitan Transportation Authority (MTA[4]). The GIS data of bike racks were obtained from the Department of Transportation (DOT[5]). The numbers of bus and subway stations and bike racks were calculated for each census tract using spatial tools of the software ArcGIS. The road network data were obtained from the LION Single Line Street Base Map sponsored by NYCDCP[6]. Road density was computed for each census tract.

MapReduce Programing

MapReduce program is used to count trips in each census tract in each month from 2010 to 2013. Pick-up trips and drop-off trips are counted separately. The input data for our mappers are unzipped from https://uofi.app.box.com/NYCtaxidata. The outputs of our MapReduce jobs consist of key-value pairs with key = census tract geoid + year_month (e.g. 201212 meaning Dec 2012), and value = count of trips. We have two output files, one for pick-up and the other for drop-off, which are aggregated basing on census tract. Thus, each file includes 4 attributes that are census tract geoid, year_month, pickup_count and dropoff_count. Reproducible steps of running MapReduce job are attached in the appendix.

During the implementation of MapReduce program, we encountered a difficulty that the coordinate system for geological location in the taxi trip files was different from the system in census tract shapefile we downloaded from nyc.gov. In order to align coordinates from trip data with geological boundaries in census tract shapefile, we created a new shapefile that contains all NYC census tracts except water area by using ArcGIS.

Rtree is a tree data structure used for spatial searching, which is proposed by Antonin Guttman in 1984. The idea behind it is to use the bounding boxes to decide whether or not to search inside a subtree. Thus, most of the nodes in the tree will never be read during a search. The below is an example of Rtree for 2D rectangles. The tree structure is to group nearby objects and represent

[4] Source: http://web.mta.info/developers/download.html
[5] Source: http://www.nyc.gov/html/dot/html/about/datafeeds.shtml
[6] Source: http://www.nyc.gov/html/dcp/html/bytes/dwnlion.shtml

them with minimum bounding rectangle in the next higher level; at the leaf level, each rectangle describes a single object and at higher levels, the aggregation of an increasing number of objects. For objects lying within a bounding rectangle, a query that does not intersect the bounding rectangle also cannot intersect any of the contained objects. To improve performance, Rtree indexing is implemented in our mapper. It has helped us reducing the time for computation tremendously.

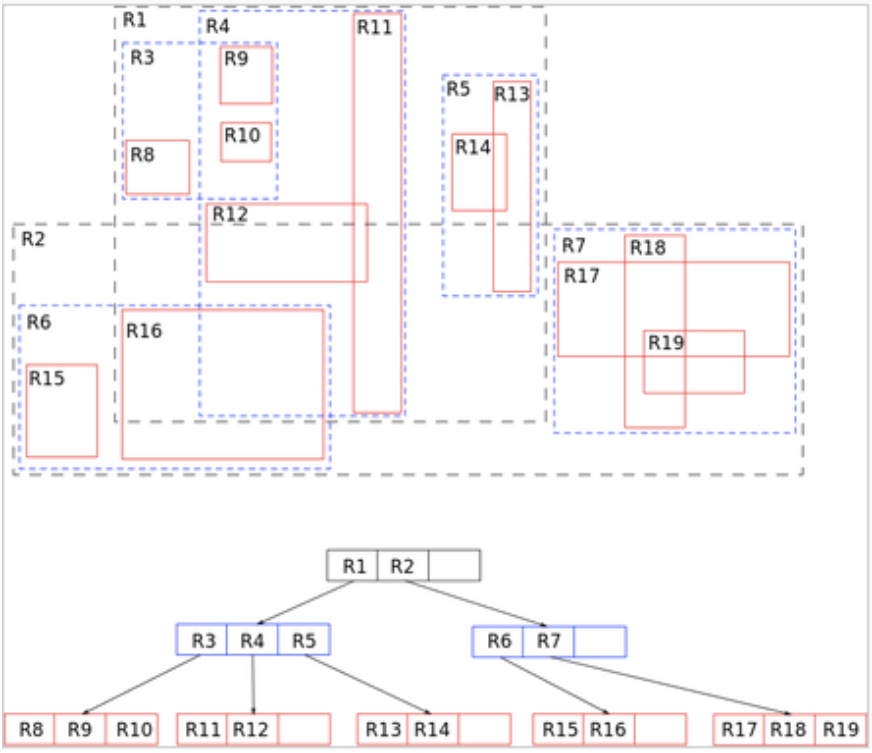

Figure 1 Rtree for 2D Rectangles

STATISTICAL MODELING

To account for the spatial dependence of taxi usage, it is assumed that the pick-ups and drop-offs of census tracts are correlated within the same neighborhood tabulation area (NTA). Therefore, the data structure used in this study can be viewed as a two-level hierarchy with level 1 being the census tract level, and level 2 being the NTA level. Hierarchical models (Gelman and Hill, 2007) that allow parameters to vary can accommodate the heterogeneity among different groups and are able to make more reliable estimations than traditional generalized linear models (GLMs). A hierarchical model is proposed as follows:

$$\ln(\lambda_{ijm}) = \beta_0 + \sum_{p=1}^{P} \beta_p X_{pijm} + T_m + S_j + \varepsilon_{ijm} \quad (1)$$

λ_{ij} denotes the pick-ups or drop-offs at ith census tract in jth NTA for mth month. X_{pijm} are explanatory variables such as median income and road density. β_0 and β_p are the regression coefficients to be estimated. It should be noted that in the hierarchical model framework β_0 and β_p can be allowed to vary across groups. T_m is month-specific temporal effect term which could account for the temporal correlation in the taxi usage. S_j is spatial effect term which is assumed to be normally distributed among the NTAs with a mean 0 and variance σ_S^2, ε_{ijm} is the model error tem which follows a normal distribution.

All model parameters were estimated using the Bayesian method that combines prior distributions with a likelihood function obtained from the observed data to estimate posterior distributions. The reason of using Bayesian method is that it has advantages in accommodating complex model structures when compared to classical statistical methods (Mitra and Washington 2007, Xie *et al.* 2013). Bayesian inference is usually implemented by a Markov Chain Monte Carlo (MCMC) algorithm (Gilks *et al.* 1998). MCMC is a classic method that utilizes independent and identically distributed simulations of a random process to approximate the desired distribution. The statistic package MCMCpack in R was used to provide a computing approach for the calibration of Bayesian models.

Bayesian posterior estimates of the pick-up and drop-off models are presented in Table 1 and Table 2. The 95% Bayesian Credible Interval (2.5% BCI, 97.5% BCI) was used to examine the significance of estimates. Estimates can be regarded as significant at the 95% level if the BCIs

do not cover 0 and vice versa (Gelman 2004). The coefficients of Feb, Jul, Aug and Oct in the pick-up model were the only four not found to be significant.

Table 1 Estimation Results of the Taxi Pick-Up Model

	Mean	**SD**	**2.5% BCI**	**97.5% BCI**
Intercept	2.9800	0.2026	2.5820	3.3810
Bus station number	0.0558	0.0008	0.0542	0.0574
Subway station number	0.2211	0.0085	0.2043	0.2376
Bike rack number	0.0254	0.0007	0.0241	0.0267
Road density (mile/mile^2)	0.0064	0.0004	0.0056	0.0072
Median income (10^3)	-	-	-	-
Employment (10^3)	0.1671	0.0054	0.1565	0.1777
Population (10^3)	-	-	-	-
Month				
Jan	0	-	-	-
Feb	-0.0315	0.0184	-0.0675	0.0046
Mar	0.0968	0.0184	0.0612	0.1331
Apr	0.0410	0.0184	0.0051	0.0771
May	0.1074	0.0184	0.0716	0.1439
Jun	0.0502	0.0183	0.0146	0.0860
Jul	0.0212	0.0183	-0.0141	0.0574
Aug	0.0169	0.0184	-0.0188	0.0529
Sep	0.1570	0.0184	0.1215	0.1933
Oct	-0.0330	0.0183	-0.0686	0.0030
Nov	-0.1098	0.0184	-0.1456	-0.0736
Dec	-0.0783	0.0184	-0.1145	-0.0420
Spatial effect σ_S^2	1.1400	0.0053	1.1300	1.1510

Table 2 Estimation Results of the Taxi Drop-Off Model

	Mean	SD	2.5% BCI	97.5% BCI
Intercept	4.8460	0.0784	4.6920	5.0000
Bus station number	0.0381	0.0002	0.0376	0.0385
Subway station number	0.1283	0.0022	0.1240	0.1326
Bike rack number	0.0111	0.0002	0.0107	0.0114
Road density (mile/mile^2)	0.0019	0.0001	0.0017	0.0021
Median income (10^3)	0.0027	0.0001	0.0026	0.0029
Employment (10^3)	0.1792	0.0029	0.1736	0.1849
Population (10^3)	0.0314	0.0013	0.0287	0.0339
Month				
Jan	0	-	-	-
Feb	-0.0520	0.0047	-0.0612	-0.0428
Mar	0.1003	0.0047	0.0911	0.1095
Apr	0.0925	0.0047	0.0832	0.1017
May	0.1291	0.0047	0.1199	0.1383
Jun	0.0915	0.0047	0.0822	0.1006
Jul	0.0841	0.0047	0.0749	0.0934
Aug	0.0649	0.0047	0.0556	0.0741
Sep	0.1171	0.0047	0.1078	0.1264
Oct	0.0426	0.0047	0.0333	0.0516
Nov	-0.0121	0.0047	-0.0212	-0.0028
Dec	0.0429	0.0047	0.0337	0.0521
Spatial effect σ_S^2	0.2933	0.0014	0.2907	0.2960

According to Table 1 and Table 2, the bus station number, subway station number, bike rack number and road density are positively associated with both the taxi pick-ups and drop-offs. It implies that taxi usage is greater in areas with higher transportation accessibility. Those areas usually have more attraction that can lead to higher traffic demand, and a portion of travelers would like to choose taxis even if the public transit is convenient. Additionally, areas with more employees are accompanied with higher taxi pick-ups and drop-offs. Median income and

population are found to have positive impacts on taxi drop-offs, but are not significant contributing factors to taxi pick-ups. The significance of the spatial effect variance σ_S^2 in both the pick-up and drop-off models confirmed the spatial dependence of taxi usage.

SPATIO-TEMPORAL ANALYSIS

Temporal Patterns of Taxi Usage

Figure 2 and Figure 3 are bar plots for the number of pick-up and drop-off in each month from 2010 to 2013. According to the plots, the number of pick-up and the number of drop-off in each month share a same pattern through these four years except for a minor difference between the amounts. The reason for such difference might is that some of the drop-off locations are not in the area of New York City. We also observed that there exists a one-year period in both pick-up number and drop-off number. In each year, the amount of pick-up/drop-off reached a peak in March, Mar or September that followed by a deceasing trend. Except for 2012, each year reached a local minimum at August and increased abruptly at September.

Figure 4 and Figure 5 are boxplots for the number of pick-up and drop-off of census tracts in each month from 2010 to 2013. In Figure 4, we could observe an obvious increasing trend for the variances of the pick-up number across census tracts through the months, which means that the pick up locations have been concentrated to some of the census tract since the total amount of the pick-up is stable. In Figure 5, the variances of the drop-off number across census tracts through the months are stable, which means the geological distribution of the drop-off locations roughly stayed the same through these four years.

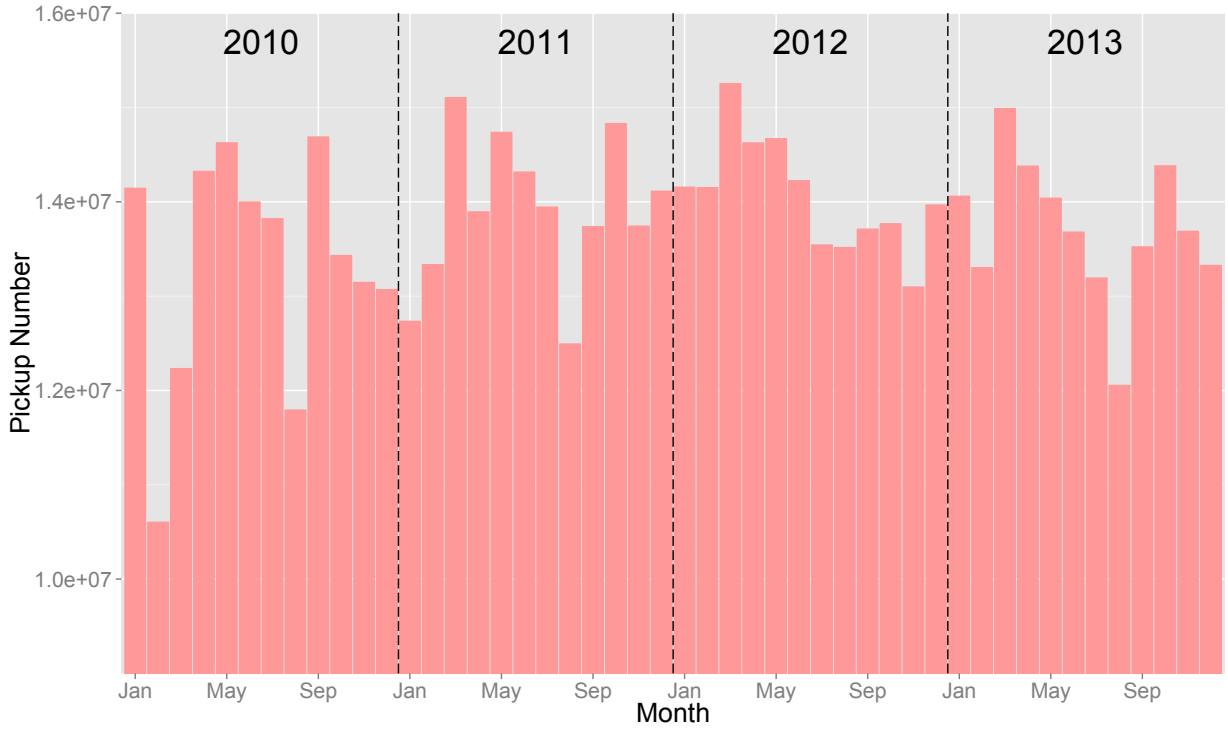

Figure 2 Monthly Taxi Pick-Up Number from 2010 to 2013

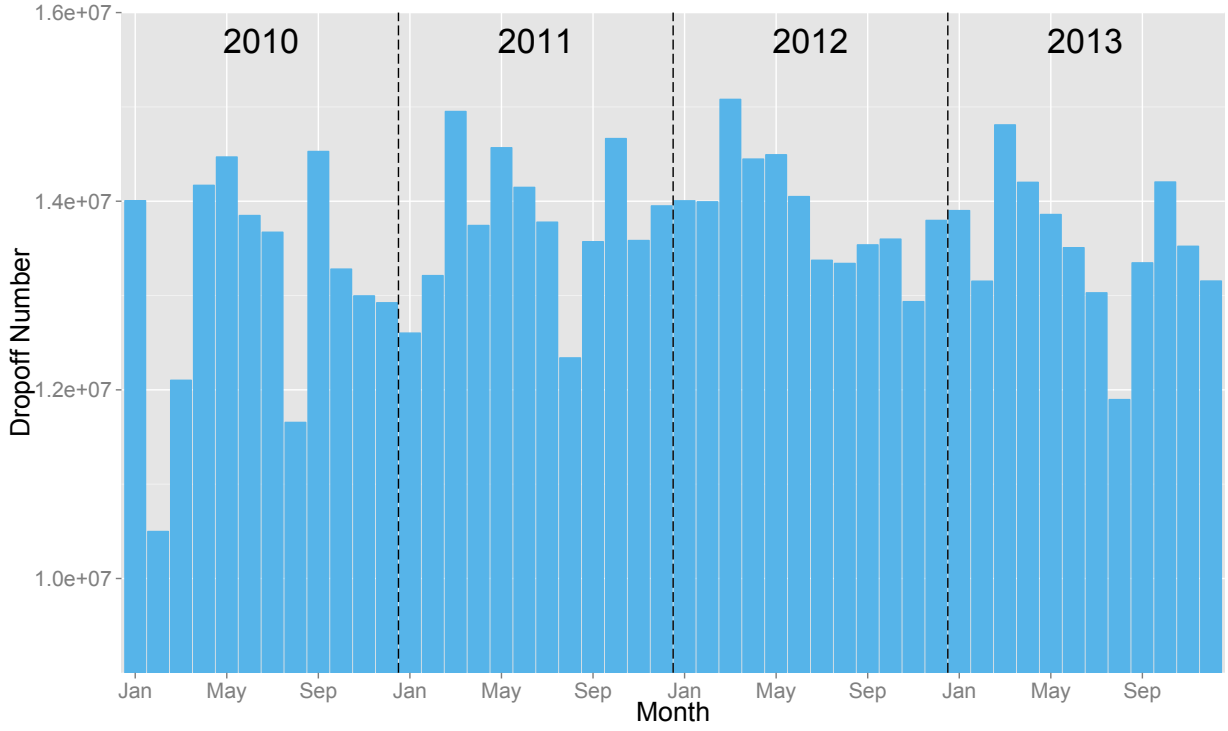

Figure 1 Monthly Taxi Drop-Off Number from 2010 to 2013

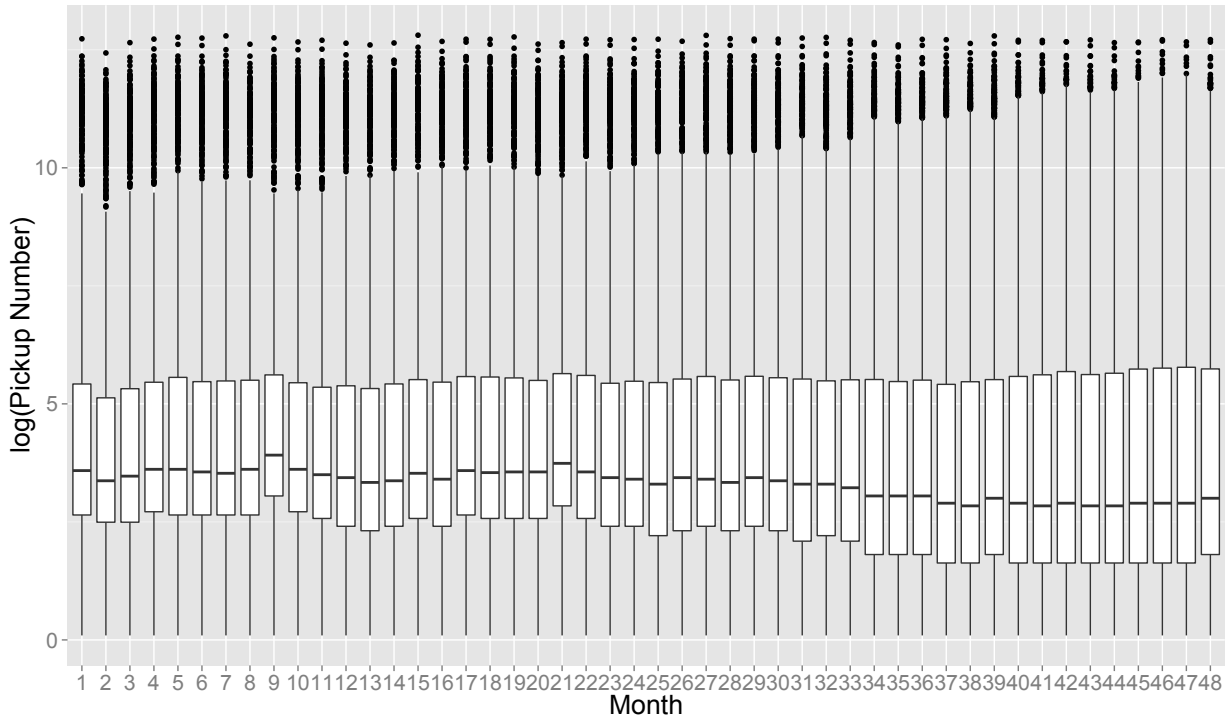

Figure 4 Boxplot for Taxi Pick-Up Number of Census Tracts from 2010 to 2013

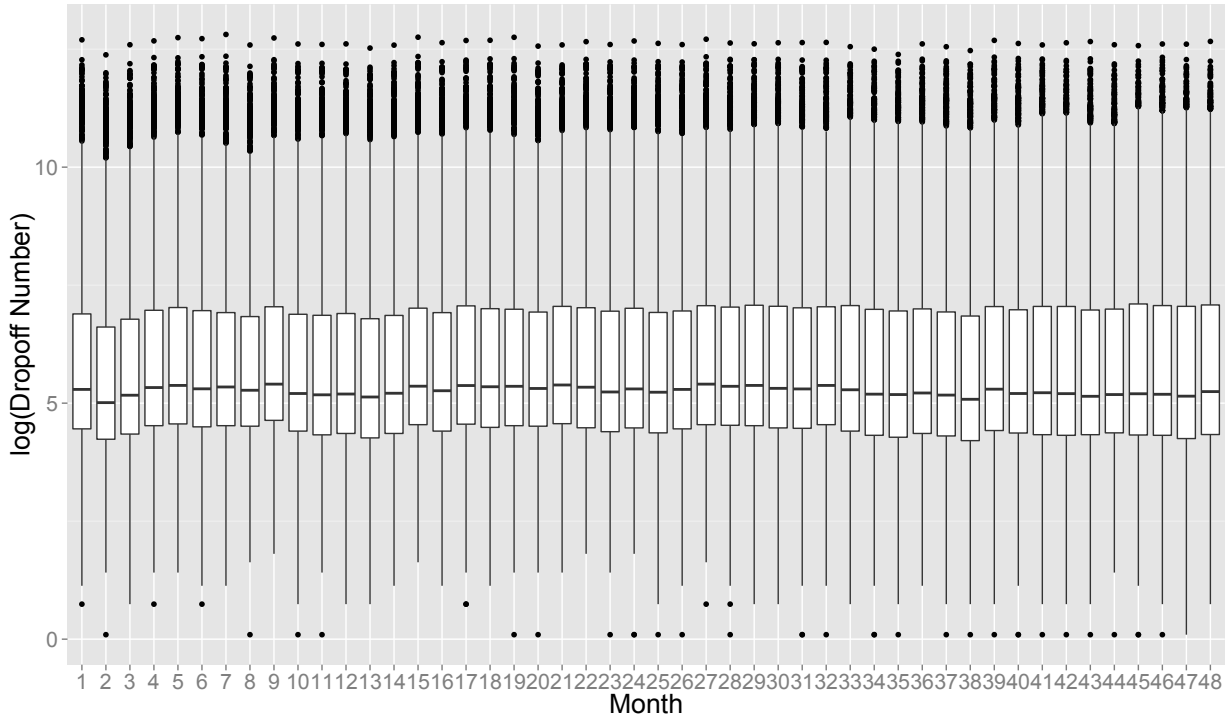

Figure 5 Boxplot for Taxi Drop-Off Number of Census Tracts from 2010 to 2013

According to coefficients of month in Table 1 and Table 2, the percentage difference of taxi pick-ups and drop-offs can be estimated using January as the base month. The confidence intervals of drop-offs are smaller compared with those of pick-ups, so the statistic inferences from the drop-off model are more reliable. The estimated percentage difference, upper bound and lower bound of 95% confidence interval are presented in Figure 6 and Figure 7. September, May and March are expected to be the top three months that have the highest pick-ups as well as drop-offs; while the February and November are predicted to among the months that have lowest taxi usage.

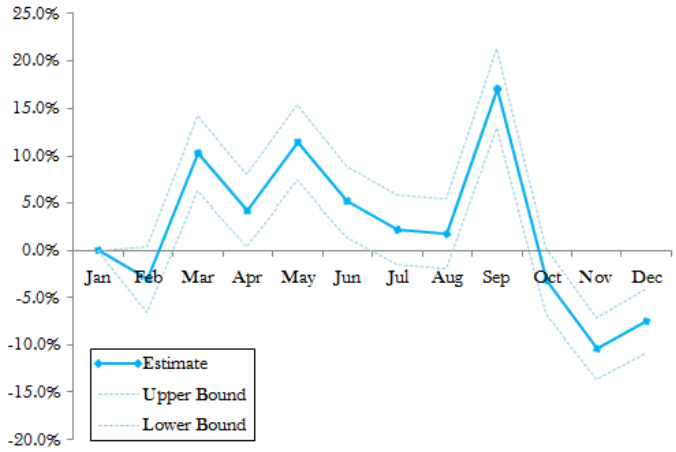

Figure 6 Estimated Percentage Differences of Taxi Pick-ups for Each Month (Use January as the Base Month)

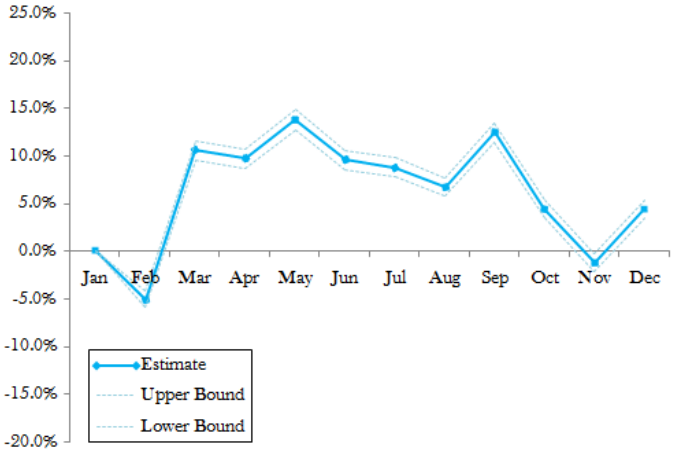

Figure 7 Estimated Percentage Differences of Taxi Drop-offs for Each Month (Use January as the Base Month)

Spatial Patterns of Taxi Usage

The total pick-ups and drop-offs from 2010 to 2013 are aggregated based on the NTA level. As presented in Figure 8, the distributions of taxi pick-ups and drop-offs are quite similar to one another. Spatial clustering can be observed in the taxi usage. The downtown and midtown Manhattan areas have significantly higher taxi usage than others. In the suburban areas, the JKF and LGA airports generate and attract a massive amount of taxi trips. The taxi usage in Staten Island and Bronx are significantly lower. The top three NTAs with the highest pick-ups are Midtown South, Hudson Yard and West Village, while the top three NTAs with the highest drop-offs are Midtown South, Hudson Yard and Upper East Side.

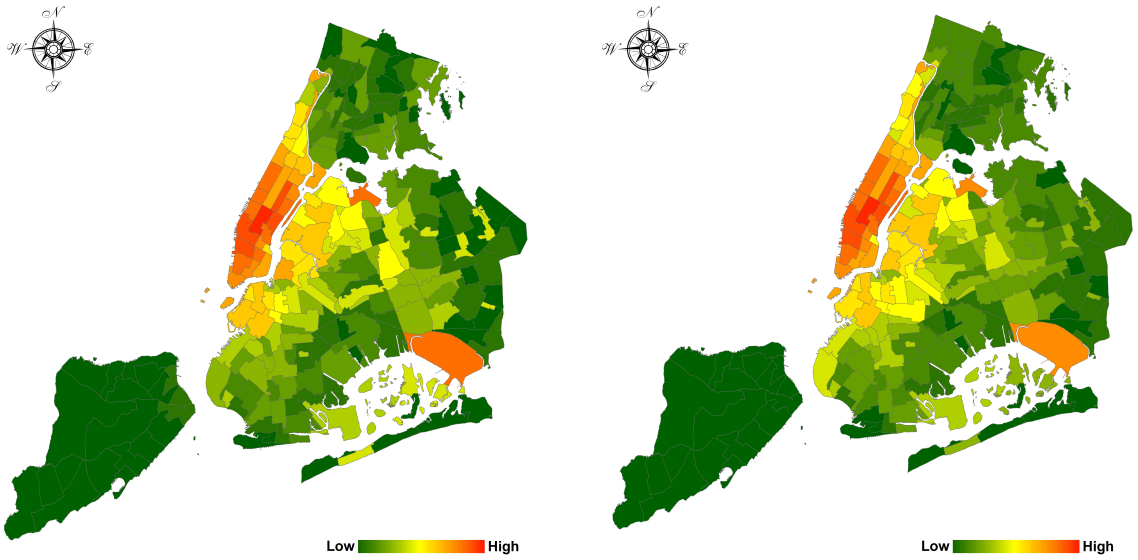

Figure 8 Observed Taxi Pick-ups (Left) and Drop-offs (Right) of Neighborhood Tabulation Areas

The NTA-specific spatial effects S_j can capture the amount of taxi trips that are contributed by unobserved factors. Maps of taxi pick-ups and drop-offs caused by unknown features are presented in Figure 9. Slight difference can be observed between Figure 8 and Figure 9. The top three NTAs with the highest "unknown" pick-ups are Airport, Turtle Bay and Upper East Side, while the top three NTAs with the highest "unknown" drop-offs are Midtown South, Airport and Turtle Bay. The reason that Airport is identified among the top three is that the taxi trips to

Airport are not derived by the explanatory variables included in the model such as high road density or employment but by some unobserved factors. Those factors can be the demand to go to airport faster or the need to carry large luggage. Examining the areas with high "unknown" trips can help better understand the taxi usage in the city.

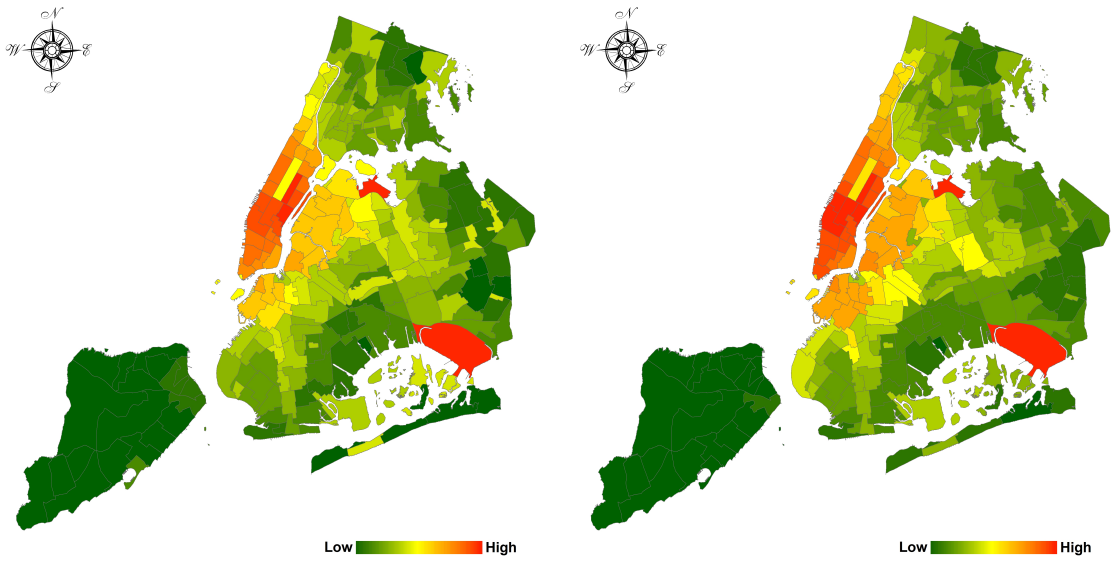

Figure 9 Taxi Pick-ups (Left) and Drop-offs (Right) of Neighborhood Tabulation Areas Caused by Unobserved Factors

SUMMARY & FUTURE WORK

This study investigates the contributing factors to taxi usage and the spatial and temporal patterns of taxi trips. A MapReduce program was designed to process the large amount of taxi data from 2010 to 2013. Monthly taxi pick-ups and drop-offs were computed for each census tract in New York City. Bayesian hierarchical models with spatial and temporal effect terms were used to identify the factors that significantly affect pick-ups and drop-offs. Seven explanatory variables including Bus station number, subway station number, bike rack number, road density, median income, employment, and population are found to affect taxi usage significantly. The difference of taxi usage among months is affirmed. September, May and March are expected to be the top three months that have the highest taxi usage; while the February and November are predicted to among those which have lowest taxi usage. Additionally, the spatial effects of taxi trips were investigated and the unobserved heterogeneity among NTAs has been confirmed.

For the future work, a more efficient way of running Bayesian models is needed. Running Bayesian Model with relatively large sample size could be extremely time consuming. Running Bayesian Model with relatively large sample size could be time consuming, since the MCMC methods, which are a set of widely used algorithms in Bayesian inference, might need long time to make the model converge. We tried to set up Hadoop to run multiple chains at the same time. However, the time we saved is limited. There are ongoing efforts to implement MCMC method in distributed and parallel setting that Mert Terzihan have tried to adapt the PyMC framework to Spark in order to run multiple MCMC chains on distributed data, while keeping PyMC's convenient abstractions for computing on probabilistic graphical models (Terzihan 2014). The reason for using Spark is that comparing to Hadoop, Spark can keep the data in distributed memory. Thus, since MCMC methods consist of iterative computations, Spark would be the most suitable platform for running MCMC jobs in parallel. Therefore, our future work after this project would be running our Bayesian model on Spark by following the instructions in Mert Terzihan's post of Bayesian Machine Learning on Apache Spark in his blog.

REFERENCES

Gelman, A., 2004. Bayesian data analysis, 2nd ed. Chapman & Hall/CRC, Boca Raton, Fla.

Gilks, W.R., Richardson, S., Spiegelhalter, D.J., 1998. Markov chain monte carlo in practice Chapman & Hall, Boca Raton, Fla.

Mitra, S., Washington, S., 2007. On the nature of over-dispersion in motor vehicle crash prediction models. Accident Analysis and Prevention 39 (3), 459-468.

Terzihan, M., 2014. Bayesian machine learning on apache spark Cloudera Engineering Blog.

Xie, K., Wang, X., Huang, H., Chen, X., 2013. Corridor-level signalized intersection safety analysis in shanghai, china using bayesian hierarchical models. Accident Analysis and Prevention 50, 25-33.

APPENDIX

Running MapReduce Program on AWS

Step 1

Create EMR cluster with following configurations

Termination protection: yes

Logging: Enabled

Hadoop distribution: Amazon AMI version 3.3.1

Roles configuration: Proceed without roles

Bootstrap Actions: click 'Add bootstrap action', choose 'Custom action', click 'Configure and add' and put the following: s3://nyubd2015liu/rtree.sh

Don't add any step at this point

Auto-termination: No

Step 2

Upload the following files to s3 bucket

Upload rtree.sh to s3 (e.g. s3://nyubd2015liu)

Upload the followings to s3 (e.g. s3://nyubd2015liu/censustract)

- mapper.py
- reducer.py
- shapefile.py
- NYC_Census_Tract.dbf
- NYC_Census_Tract.prj
- NYC_Census_Tract.sbn
- NYC_Census_Tract.sbx
- NYC_Census_Tract.shp
- NYC_Census_Tract.shp.xml
- NYC_Census_Tract.sbx

Unzip and upload 4-year trip data to s3 (e.g. s3://nyubd2015liu/finalproject/inputdata)

Note:

Rtree.sh located at https://github.com/ViDA-NYU/aws_taxi

Source codes of mapper and reducer, shapefile.py and shapefile data at: https://github.com/coolshaker/taxi_project/tree/master/source_censustract_pickup

4-year taxi trip data at https://uofi.app.box.com/NYCtaxidata

Step 3

Add the following streaming configurations to hadoop job

Mapper: s3://nyubd2015liu/censustract/mapper.py

Reducer: s3://nyubd2015liu/censustract/reducer.py

Input: s3://nyubd2015liu/tax_project2015/inputdata/

Output: s3://nyubd2015liu/output

Arguments: -D mapred.reduce.tasks=1 -files s3://nyubd2015liu/censustract/mapper.py,s3://nyubd2015liu/censustract/reducer.py,s3://nyubd2015liu/censustract/shapefile.py,s3://nyubd2015liu/censustract/NYC_Census_Tract.shp,s3://nyubd2015liu/censustract/NYC_Census_Tract.shp.xml,s3://nyubd2015liu/censustract/NYC_Census_Tract.dbf,s3://nyubd2015liu/censustract/NYC_Census_Tract.prj,s3://nyubd2015liu/censustract/NYC_Census_Tract.sbn,s3://nyubd2015liu/censustract/NYC_Census_Tract.sbx,s3://nyubd2015liu/censustract/NYC_Census_Tract.shx

Step 4

Wait for the job to finish and download output file of pickup trip counts.

Step 5

Modify lng_id from 10 to 12, lat_id from 11 to 13 in mapper.py. Run above steps 1 and 3 one more time to generate drop-off trip counts, and then download the output file.

Contributors

Kun Xie

- Collect and process public transit, road network data
- Develop Bayesian hierarchical models with spatial and temporal effects
- Visualize the temporal and spatial patterns of taxi usage
- Write the report section “Abstract”, “Introduction”, “Data Sources”, “Statistical Modeling”, “Temporal Analysis” and “Summary & Future Work”

Yuzheng Zhuang

- Collect demographic, and socioeconomic data for each census tract
- Debug MapReduce program on AWS
- Visualize the temporal patterns of taxi usage
- Write portions of report section “Temporal Patterns” and “Future Work”

Ya Liu

- Write MapReduce program for taxi usage calculation
- Run MapReduce program on AWS
- Write the report section “MapReduce Programing”

GitHub Repository

https://github.com/coolshaker/taxi_project