Predicting Taxi Demand-Supply Mismatches to Dynamically Position Mobility-on-Demand Services

Final Report

Shashi Shekhar Jha Menusha Milaj Shih-Fen Cheng Archan Misra
Singapore Management University
{shashij,menusham,sfcheng,archanm}@smu.edu.sg

ABSTRACT

Examining people's travel patterns is essential for urban planning and efficient transportation system. In this article, we aim to develop a solution framework to mitigate the taxi demand and supply imbalances across the city by utilizing the mobility patterns of commuters in densely populated cities like Singapore and taxi availability data. We then propose a mechanism to predict the spatio-temporal distribution of commuter demand in real-time. This information can be used to dynamically re-position the mobility-on-demand (MOD) vehicles (such as taxis and self-driving cars) to smooth demand-supply imbalances.

ACM Reference format:
Shashi Shekhar Jha Menusha Milaj Shih-Fen Cheng Archan Misra. 2017. Predicting Taxi Demand-Supply Mismatches to Dynamically Position Mobility-on-Demand Services. In *Proceedings of DataSpark Mobility Open-Task Challenge, Singapore, November 2017 (CIKM-DataSpark Challenge'17),* 4 pages.
DOI:

1 PROBLEM STATEMENT

In most cities, taxis play an important role in providing point-to-point transportation service. If the taxi service can be made to be reliable, responsive, and cost-effective, past studies show that taxi-like services (in the literature, such services are usually termed as MoD, *mobility-on-demand*, or MaaS, *mobility-as-a-service*) can be a viable choice in replacing a significant amount of private cars [4, 7, 8]. However, as pointed out by earlier studies [1], making taxi services efficient is extremely challenging, mainly due to the fact that taxi drivers are self-interested and they operate with only local information. A critical first step in improving taxi services is thus to provide credible ways in predicting demands for taxi services.

In this report, we discuss our solution to create a scalable and effective demand prediction system for the taxi services. Our proposed demand prediction system is designed to be operating in real-time and incorporating multiple data sources. This is in contrast to the prior studies that has focused on using only historical taxi pickup data [5, 6, 9]. The major issue with these past approaches is the fact that they focus only on *realized* demands and neglect *unrealized* demands. By incorporating information derived from mobile-phone traces, we can improve the accuracy of our demand prediction engine.

This engine can serve as the foundation for a city-scale driver guidance system (DGS) that could significantly improve the performance of the taxi fleet. We argue that such system is of critical importance in helping taxi operators surviving intense competitions from new entrants such as Uber and Grab (a major ride-hailing firm that offers Uber-like services in the Southeast Asia).

Our main contributions are:

- We describe how a driver guidance system can be designed and implemented to significantly improve the performance of taxi fleets.
- We design and implement a real-time demand prediction engine at the street level. Compared to the state-of-the-art approach from the literature, we show that our approach performs significantly better.
- We demonstrate how mobility data derived from mobile phones can be used to further improve the prediction quality of our demand prediction engine.

2 DATA SOURCES

For our demand prediction model, we use the following data sources which include all three data sources from DataSpark and one publicly available dataset on taxi availability:

Source 1: Footfall. This data source provides the estimated number of users in a particular region at a given time window.

Source 2: Origin-destination flow. This is our main source of data, that helps us to understand the flows of user movements between regions. By carefully analyzing the ingress and egress flow of users in a region, we estimate the distribution of the resident users in space and time.

Source 3: Dwell time. This data source enables us to understand users' stay patterns (i.e., residency time) in a region.

Source 4: Taxi availability. While the above mentioned data sources provide hourly statistics of user mobility patterns, taxi availability data[1] gives finer-grained details of free taxis (e.g., GPS locations along with timestamps). This provides information on real-time taxi supply.

CIKM-DataSpark Challenge'17, Singapore
© 2017 ACM. …$15.00
DOI:

[1] https://data.gov.sg/dataset/taxi-availability

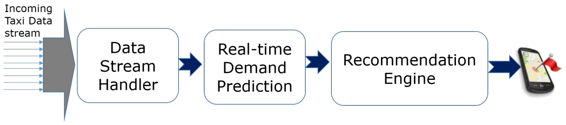


Figure 1: System Architecture

3 THE DESIGN OF THE DRIVER GUIDANCE SYSTEM

Taxi industry is diverse and complicated and is structured and managed very differently from city to city. To put this report in context, we build and test our system using Singapore as the testbed. Nonetheless, we believe our framework is general enough to be used in other cities; however, some components might need to be modified or re-calibrated depending on the operating environment.

For our study, we have collaborated with the Land Transport Authority (LTA) of Singapore, which oversees all aspects of land transportation including taxi matters, and acquired the taxi demand data for multiple months. This data provide us the realized demand information for each subzone and street of Singapore.

The design of our framework is highlighted in Figure 1. There are three most important components: 1) the handling of incoming stream data, 2) real-time demand predictions, and 3) driver's recommendation engine.

To support real-time decision support, we design our platform to accept streaming data, assuming that every 30 seconds, up to 26,000 state updates will be coming in through a private API. To support real-time sensing of both taxi locations and to support demand predictions, we have to continuously update the locations of all unique taxis based on received data. There are two major practical difficulties:

- Incoming data can contain errors for a number of different reasons. To ensure that our engine is free of apparent errors, we have to monitor all potential exceptions and handle them appropriately.
- To make best use of taxi coordinates, we need to map these coordinates to the actual road links.

In the interest of space, we will not go into the details of the engineering designs we came up in tackling these challenges. Instead, we will focus on the design of taxi demand prediction engine.

4 THE DEMAND PREDICTION ENGINE

The design of our demand prediction engine is two tier viz. 1) Street level and 2) Subzone level. Singapore is divided in a total of 323 subzones. Figures 4 shows the outlines of all the subzones boundaries over the map of Singapore. We design the demand prediction model at the street level to provide the likelihood of finding a passenger on a street within a subzone while the subzone level demand prediction helps in balancing the demand and supply across the whole Singapore.

4.1 Street Level Demand Prediction

The street level demand prediction model focuses mainly on demand generation potential for each individual street. The key insight we utilize is to treat each free-cruising taxi as a *demand probe.* The chance of us seeing demand on a particular link is assumed to be inversely correlated with the amount of time passed since last visit by a free taxi. In other words, we assume that the demand arrivals do not follow a Poisson process, as the memoryless property is not valid.

Formally speaking, for each cruising taxi approaching a street, the time elapsed since the latest arrival of a taxi having AVAILABLE status on that same street is maintained. According to the model, this elapsed time acts as the independent variable while the status of the approaching taxi when it exits that street (which could be HIRED or NOT-HIRED) acts as the dependent variable. The model then uses these two variables to run a multilevel logistic regression [2] with grouping based on street, time of the day and day of the week. For the time of the day, a period of 30 minutes is considered as one time-slot resulting in a total of 48 time-slots in a day.

The multilevel regression model returns the likelihood of getting a passenger on a street given the elapsed time from the latest taxi in the AVAILABLE state which approached the street. The following equation describes this prediction model:

$$\Pr(\text{HIRED}|\delta_s) = \text{logit}^{-1}(\alpha_{s,t,d} + \beta_{s,t,d}\delta_s), \tag{1}$$

where s is the street, t is the time slot of the day and d is the day of the week. α and β are the coefficients of the regression model while δ_s is the elapsed time from the latest arrival of a taxi in the AVAILABLE state on the street s. $\Pr(\text{HIRED}|\delta_s)$ signifies the probability of getting a passenger for a cruising taxi.

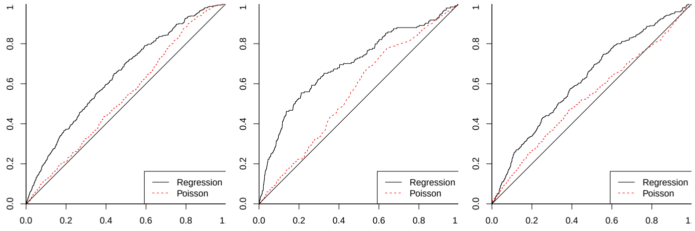
For evaluation of the predictive capability of this street level demand prediction model, we compared it against a non-homogeneous Poisson model [3]. The non-homogeneous Poisson model has a time dependent rate function ($\lambda(t)$). Since the demand of taxis varies with the time of the day, a time dependent rate function with a cycle of 24 hours (48 time slots in our case) is adopted as a piecewise linear function. The following equation describes the Poisson model with time dependent rate function $\lambda(t)$:

$$\Pr(t_s) = \lambda(t)e^{-\lambda(t).t_s}, \tag{2}$$

where t_s is the time from the last trip (street hail) on the street s and t is the current time-slot.

The graphs in Figure 2 show the ROC curve of three different streets for using the predicted demand values of a single day. The ROC curve plots the True positive rate (Sensitivity) versus the False positive rate (Specificity) for a predictor at different threshold settings. The black colored curve in the graph shows the ROC for the regression model while the red curve is for the Poisson model. The diagonal line ($x = y$) bisecting the graph in two equal halves is the line of random guess. Hence, the points above the diagonal line indicate a better prediction characteristic. As can be observed from the graphs, the regression based predictor outperforms the Poisson based predictor.

The graph in Figure 3 depicts the Area Under the Curve (AUCs) of ROCs of different streets in Singapore for both the prediction models. We discovered from the historical data that a set of streets account for almost 70% of street pickups across several months. The data from these streets are used for generating output.



(a) (b) (c)

Figure 2: Comparison of ROC curves of the logistic regression model against the non-homogeneous Poisson model for the prediction of the likelihood of demand for three different streets. The X-axis shows the false-positive rate while the Y-axis is for the true-positive rate.

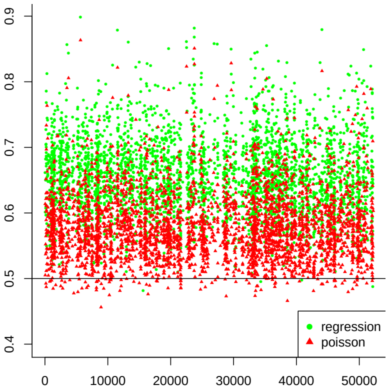


Figure 3: Comparison of AUCs for the regression based demand prediction model in the framework versus Non-Homogeneous Poisson model. The AUCs are shown on the Y-axis while the X-axis lists the street ids of the streets considered for evaluation.

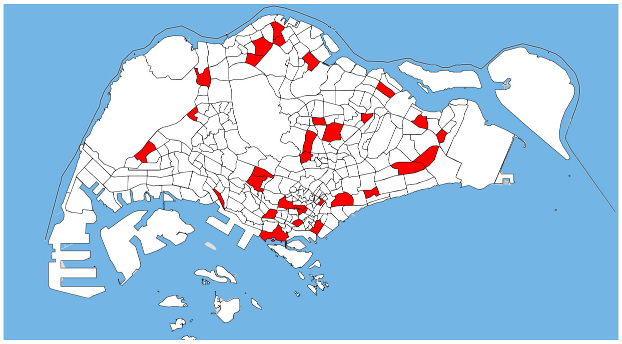


Figure 4: The map of Singapore with subzone boundaries.

4.2 Subzone Level Demand Prediction

The first part of our demand prediction engine, as described in the previous section, is based on the insight that demands from a large number of road links do not follow Poisson arrival process.

To predict the expected demand count at the subzone level, we utilize the mobility-related data provided by the DataSpark API along with the real-time Taxi Availability data from a publicly available API in Singapore. Due to the restricted API access, we selected a sample of 31 subzones to study the prediction quality. These 31 subzones are highlighted (in red) in Figure 4. The demand (trips) in the selected sample of subzones varied from 223,374 to 6434 trips in the Month of May 2017.

We have used OLS linear regression to model the demand prediction at the subzone level. Since the mobility-related data from DataSpark is available on hourly basis, we have used the time period of one hour to aggregate the trips in each subzone from the month of May 2017. Further, the taxi availability data is averaged for each hour of the day. The OLS models are depicted in the following equations.

$$
\begin{aligned}
M1:\ y_i &= \beta_0 + \beta_1 T_a, && (3)\\
M2:\ y_i &= \beta_0 + \beta_1 O_g, && (4)\\
M3:\ y_i &= \beta_0 + \beta_1 T_a + \beta_2 O_g, && (5)\\
M4:\ y_i &= \beta_0 + \beta_1 T_a + \beta_2 O_g + \beta_3 I_n + \beta_4 D_t + \beta_5 F_t && (6)
\end{aligned}
$$

where y_i is the predicted demand for a subzone i at a particular time of the day, $\beta_0 - \beta_5$ are model coefficients, T_a is the taxi availability, O_g is the egress from the subzone i, I_n is the ingress in the subzone i, $D_t$is the average dwell time and F_t is the ratio of unique and complete footfall within the subzone i.

The quality of the subzone level demand prediction model is evaluated using three quality metrics viz. symmetric Mean Absolute Percentage Error (sMAPE), Akaike Information Criterion (AIC) and Root Mean Square Error (RMSE). The model with least sMAPE, AIC and RMSE is selected. Table 1 shows the average of different performance metrics aggregated over all the 31 subzones with different independent input variables. We used May 2017 data for deriving model coefficients. As the dependent variable, we used the trips in the succeeding hours against the input from the current hour to make the prediction practically viable for real use.

As can be observed, the base model (M1) using only the Taxi Availability data has almost 43% aggregated average sMAPE error while the combined model of Taxi Availability and Egress (M3) from subzone reduces the sMAPE error by almost 5% with the reduction in RMSE at almost 13%. The Full Model (M4) uses a set of 6 inputs within a subzone at a time. As can be noted from Table 1, in case of Full Model, the average sMAPE reduces by 10% while the reduction in RMSE is around 23.5%. In addition, the AIC value of the model M4 is also the least. Further, we also performed ANOVA analysis among the different regression models. We found that the F static was significant at 5% level for all the 31 subzones for the model M4 which provides stronger evidence for the effective contribution of mobility related information in the model. Hence, we selected the Model M4 for generating the subzone level demand prediction.

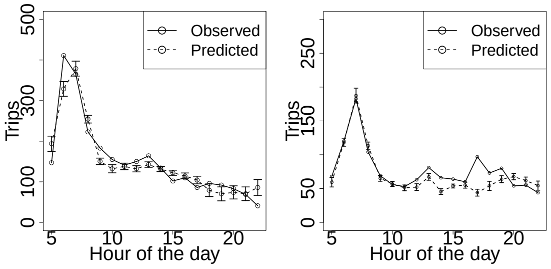
In order to evaluate the effectiveness of the subzone level demand prediction, we selected the first week of June 2017 (5-9 Jun'17) to test the prediction accuracy. Table 2 lists the sMAPE error for five different subzones from June 5-9 2017 (Monday to Friday). As can be observed, the sMAPE errors are either below or close to the model average sMAPE. Further, the graphs in Figure 5 show the

Table 1: Relative Performance of different models taking aggregated averages of all 31 subzones.

OLS Regression Model	sMAPE	AIC	RMSE
M1	0.430	4369.312	56.132
M2	0.410	4321.793	54.199
M3	0.381	4253.716	49.039
M4	0.328	4139.447	42.920

Table 2: sMAPE for the test days from June 5-9 2017 for five high demand subzones.

Day	BDSZ04	TMSZ04	SRSZ01	WDSZ03	DTSZ01
Mon	0.205	0.185	0.157	0.169	0.327
Tue	0.207	0.273	0.174	0.185	0.314
Wed	0.263	0.332	0.352	0.181	0.439
Thu	0.263	0.369	0.265	0.239	0.325
Fri	0.244	0.264	0.233	0.165	0.342



(a) Monday (WDSZ03)

(b) Friday (SBSZ03)

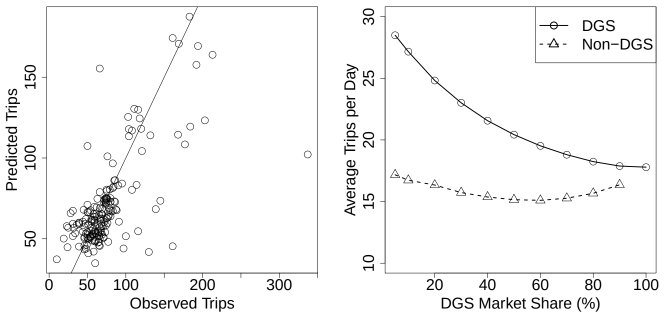
Figure 5: Predicted and observed demands for two working days in two different subzones. The X-axis shows the hour of the day while the Y-axis shows the trip count.

observed and predicted trips against the different hours of the day with confidence interval on a Monday and Friday in two different subzones. As can be observed from the graphs, the predicted trips closely approximates the observed trips on both days of the week. The graph in Figure 6a depicts the predicted versus observed trips for the whole week in the subzone WDSZ03.

The results show that the availability of mobility related information can greatly help to estimate the taxi demand at the subzone level at different times of the day.

5 APPLICATION DESIGN

Our application is mainly designed to support the operation of the taxi driver guidance system. We use the subzone level demand prediction to provide island level recommendations to the taxi drivers. This helps in alleviating the demand-supply imbalance across the city and places the taxis at the right locations. At the next level of recommendation, we guide the taxis to a particular street within a sub-zone with a high likelihood of finding passenger. To allow users visualize the prediction outcomes, we have designed a series of dashboard components for easily visualization available datasets and real-time/historical demand predictions.



(a)

(b)

Figure 6: (a) Observed versus predicted demand from 5-9 Jun'17 for the subzone WDSZ03. (b) The performance of DGS and non-DGS drivers as a function of the adoption ratio of the DGS technology.

6 IMPACT OF THE PROPOSED SOLUTION

The most important impact we expect to create with our proposed solution is the improved taxi driver guidance system (DGS). In our initial simulation study, we show that by adopting the DGS, drivers can achieve significant increase in their productivities. This benefit holds even with relative high adoption ratio. The performance of DGS drivers over non-DGS drivers, under different market adoption ratio, can be seen in Figure 6b.

REFERENCES

[1] Y. Ding, S. Liu, J. Pu, and L. M. Ni. 2013. HUNTS: A Trajectory Recommendation System for Effective and Efficient Hunting of Taxi Passengers. In *2013 IEEE 14th International Conference on Mobile Data Management*, Vol. 1. 107–116. https://doi.org/10.1109/MDM.2013.21

[2] Andrew Gelman and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models.* Cambridge university press.

[3] Jerald F Lawless. 1987. Regression methods for Poisson process data. *J. Amer. Statist. Assoc.* 82, 399 (1987), 808–815.

[4] Todd Litman. 2000. Evaluating carsharing benefits. *Transportation Research Record: Journal of the Transportation Research Board* 1702 (2000), 31–35.

[5] Xi Liu, Li Gong, Yongxi Gong, and Yu Liu. 2015. The orienteering problem: A survey. *Journal of Transport Geography* 43, 1 (2015), 78–90.

[6] Yu Lu, Gim Guan Chua, Huayu Wu, and Clement Shi Qi Ong. 2016. An Intelligent System for Taxi Service Monitoring, Analytics and Visualization. In *Twenty-fifth International Joint Conference on Artificial Intelligence (IJCAI).* 4256–4257.

[7] Shuo Ma, Yu Zheng, and Ouri Wolfson. 2015. Real-time city-scale taxi ridesharing. *IEEE Transactions on Knowledge and Data Engineering* 27, 7 (2015), 1782–1795.

[8] Dongxu Shao, Wei Wu, Shili Xiang, and Yu Lu. 2015. Estimating taxi demand-supply level using taxi trajectory data stream. In *IEEE International Conference on Data Mining Workshop (ICDMW).* IEEE, 407–413.

[9] Dongxu Shao, Wei Wu, Shili Xiang, and Yu Lu. 2015. Estimating Taxi Demand-Supply Level Using Taxi Trajectory Data Stream. In *IEEE International Conference on Data Mining Workshop (ICDMW).* 407–413.