Inferring Unmet Human Mobility Demand with Multi-Source Urban Data

Kai Zhao[1], Xinshi Zheng[1], and Huy Vo[2]

[1] Center for Urban Science and Progress,
New York University,
kai.zhao@nyu.edu, xinshi.zheng@nyu.edu
[2] Department of Computer Science,
City College of the City University of New York
huy.vo@nyu.edu

Abstract. As the sharing economy has been increasing dramatically in the world, the mobile-hailed ridesharing companies like Uber and Lyft in the US, Didi Chuxing in China has begun to challenge traditional public transportation providers such as bus, subway or taxis. Ridesharing companies have shown their ability to provide the mobility services where public transit has failed. The human mobility demand that cannot be satisfied by traditional transportation modes (unmet human mobility demand) can be served by the ridesharing companies. In this paper, we provide a 'hydrological' perspective for inferring unmet mobility demand patterns in cities with multi-source urban data. We observe that the unmet human mobility demand is proportional to the met mobility demand by examining the yellow taxi and the Uber data in New York City. Based on this observation, a Single Linear Reservoir (SLR) model has been proposed for modeling unmet human mobility demand from multi-source urban data.

Keywords: urban human mobility; spatio-temporal data mining

1 Introduction

As the sharing economy has been increasing dramatically in the world, the mobile-hailed ridesharing companies like Uber and Lyft in the US, Didi Chuxing in China has begun to challenge traditional public transportation providers such as bus, subway or taxis. Ridesharing companies have shown their ability to provide the mobility services where public transit has failed. The human mobility demand that cannot be satisfied by traditional transportation modes (unmet human mobility demand) can be served by the ridesharing companies [10].

In this paper, we provide a 'hydrological' perspective for inferring unmet mobility demand patterns (e.g., Uber) in cities with multi-source urban data (e.g., subway, taxi, bus, or bike data). We observe that the unmet human mobility demand is proportional to the met mobility demand by examining the yellow taxi and the Uber data. A SLR model [3] developed previously for rainfall-runoff

analysis has been adopted for modeling unmet mobility demand. The primary goal of this paper is to establish the mathematical relationship between the urban human mobility demand and the unmet mobility demand, and hence develop a 'unit hydrograph' methodology for predicting future unmet mobility demand.

In this paper, urban human mobility status has been simplified with the following assumptions: 1) The mobility demand can be met by any type of public transportation including taxi, bus, subway and bike; 2) The demand cannot be met for a given time period will eventually be met by ridesharing transportation modes such as Uber. The total urban human mobility demand is defined as the number of passengers in a region that have travelling demand using ground transportation, such as subway, taxi, bike or bus, during a given time interval. The met urban human mobility demand is the amount of passengers that are able to find a method of transportation within this given time period. The unmet mobility demand is the number of passengers that cannot find any transportation mode during this given time period.

First, we observe that the unmet human mobility demand is proportional to the met mobility demand. Based on this observation, we borrow the 'unit hydrograph' concept in hydrology, which is originally the unit pulse response of runoff for a watershed receiving a unit amount of excess rainfall for a given duration, is defined here as the response of the number of passengers to choose Uber given a unit input of urban human mobility. However, unlike hydrological studies, which are focused on deriving the 'unit hydrograph' to estimate the runoff from rainfall input, our aim is to use this method to estimate the change of storage within the linear reservoir when the total urban human mobility demand changes. The reason we use the SLR model is that, the urban human mobility demand [10] is similar to the rainfall-runoff [3]. It is dynamic with spatially distributed inputs and outputs. There is a peak of the human mobility demand, like the rainfall input, and a concentration time that the human mobility demand to be full-filled, like the time the water needs to pass through the system. The unmet human mobility demand, similar to the runoff water, is the amount of passengers plan to travel from one location to another, but cannot find the right public transportation system.

2 Positive Correlation Between the Taxi and Uber Demand

Ridesharing companies have shown their ability to provide the mobility services where public transit such as taxi or bus has failed. The human mobility demand that cannot be satisfied by traditional transportation modes, i.e., the unmet human mobility demand, can be served by the ridesharing companies such as Uber or Lyft. In this section, we show that the unmet human mobility demand is proportional to the met mobility demand by examining the yellow taxi and the Uber data in New York City [12]. The yellow taxi and Uber utilize different cruising strategies. The yellow taxis usually use a random cruising strategy, while the Ubers can go to the passenger's places when a request is received. Therefore,

when traditional transportation modes fail to meet the human mobility demand, ridesharing companies such as Uber have the potential to satisfy the unmet human mobility demand.

2.1 Data-sets

First, we give a description of the data-sets used in this paper and how we pre-process the data:

Taxi and Uber Data-sets The NYC yellow taxi data-set is a public data-set provided by the Taxi and Limousine Commission (TLC) [7]. TLC records the information from all trips completed in yellow taxis in NYC. Each trip record includes fields capturing pick-up and drop-off time, pick-up and drop-off locations, trip distances, itemized fares, and passenger counts. In total we have 13,813,031 taxi pick-up records from 13,237 yellow taxis.

The New York Uber data-set is a public data-set from TLC [7] that aims to study the Uber behaviour. This data-set contains 663,845 Uber pick-up records. We examine both the yellow taxi and Uber data-sets for one month (June 2014). We extract following information from the data-set: taxi ID, pick-up time and the corresponding pick-up location (neighborhood).

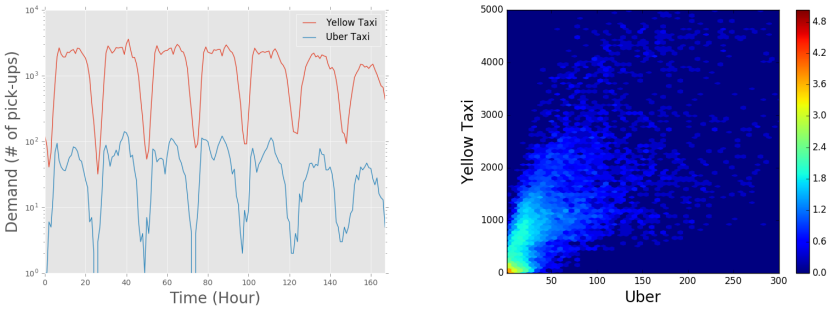
Data Pre-processing We use the NYC neighborhood (in total 184 neighborhoods) shape file [6] to map the pick-up GPS points with the associated neighborhoods: if the pick-up location is within the neighborhood, we consider that neighborhood as the one passengers getting on the taxi and there is one mobility demand at that neighborhood. All of our data pre-processing were conducted using the operational data facility at our research center [4, 5]. In particular, the mapping of taxi pick-ups to geospatial features, which requires a lot of processing given the volume of the pick-up trips, on a 1200+ core cluster running Cloudera Data Hub 5.4 with Apache Spark 1.6. The cluster consists of 20 high-end nodes, each with 24TB of disk, 256GB of RAM, and 64 AMD cores.

2.2 Linear Correlation Between Met and Unmet Human Mobility Demand

We find that there is a strong positive correlation between the yellow taxi (met mobility demand) and Uber (unmet mobility demand) pick-ups (see Fig. 1 (a)). Fig. 1 (b). shows the density of hourly pick-ups of yellow and Uber in the 184 neighborhood. A high density of points are near diagonal line, identifying a clear positive correlation.

We use the Pearson correlation coefficient to quantify the strength of the correlation between the hourly pick-ups of yellow and Ubers in all 184 neighborhoods. Our observation verifies our proof in the SLR model with a Pearson values as 0.82. The $p-$value is less than 0.01, identifying a very strong statistical significance. We show that there is a strong positive correlation between the

unmet human mobility demand served by Uber and the met mobility demand served by yellow taxi in New York City.



(a) Yellow taxi and Uber demand (b) Pearson Correlation

Fig. 1: (a) One week met mobility demand (yellow taxi) and unmet mobility demand (Uber) in June, 2014. (b) Positive correlation between met mobility demand and unmet mobility demand.

3 Linear Reservoir Model for Unmet Mobility Demand Estimation

Since we observe that the unmet human mobility demand is proportional to the met human mobility demand by examining the yellow taxi and the Uber data. In this section, we propose a Single Linear Reservoir (SLR) model for modeling unmet mobility demand based on this observation. The primary goal is to establish the mathematical relationship between the urban human mobility demand and the unmet mobility demand, and hence develop a 'unit hydrograph' methodology for predicting future unmet mobility demand.

The reason we use the SLR model is that, urban human mobility demand phenomena is similar to the rainfall-runoff pattern in hydrology. It is dynamic with spatially distributed inputs and outputs. The system is regarded as a single idealized reservoir with unmet mobility demand as storage (S). The human mobility demand D (A in discrete time scale) is the input (rainfall) into the reservoir, and the portion of mobility demand that will be met within a specified time interval will be the output (runoff) of the reservoir (M). Here the human mobility demand A can be inferred with multi-view learning algorithm [9].

The linear reservoir model will first be analyzed in continuous time scale, and then converted to discrete time scale for application in developing 'unit hydrograph' for unmet mobility demand analysis. For an ideal linear system, there is continuity equation on continuous time scale:

$$\frac{dS(t)}{dt} = D(t) - M(t) \tag{1}$$

where t is time.

Since the system is linear, it is reasonable to assume that the unmet mobility demand is proportional to the output, the met mobility demand. There is:

$$S(t) = kM(t) \tag{2}$$

where k is a constant response factor that can either be determined from historical input, output data, or from the characteristics of the studied urban region.

Combining equation (1) and equation (2), there is:

$$k\frac{dM(t)}{dt} + M(t) = D(t) \tag{3}$$

The unit impulse response of the system occurs when the system receives an input of unit amount instantaneously. Let the unit impulse response function at time t be $u(t-\tau)$ (The impulse occurs at τ). There is convolution integral from the two linear system principles of proportionality and superposition:

$$M(t) = \int_0^t D(\tau)u(t-\tau)d\tau \tag{4}$$

The unit step response of a system $r(t)$ is resulted from an input that changes from 0 to 1 at time 0 and continues indefinitely at that rate thereafter. With equation (4) $r(t)$ is found for $D(\tau) = 1$ for $\tau \geq 0$:

$$r(t) = \int_0^t u(t-\tau)d\tau = \int_0^t u(l)dl \tag{5}$$

where $l = t - \tau$.

The unit pulse response function $p(t)$, which is resulted from an input of unit amount occurring in duration Δt, can be determined based on the two linear system principles:

$$p(t) = \frac{1}{\Delta t}[r(t) - r(t-\Delta t)] = \frac{1}{\Delta t}\int_{t-\Delta t}^{t} u(l)dl \tag{6}$$

Similar to hydrological data, urban mobility data will be in discrete time intervals. In discrete time intervals of duration Δt, for the input of the system there is:

$$A_i = \int_{(i-1)\Delta t}^{i\Delta t} A(\tau)dt \tag{7}$$

where A_i is the accumulated urban human mobility demand during the time interval Δt. And $i = 1, 2, 3, ..., I$, where I is the last time interval of Δt. $D(\tau) = D_i/\Delta t$ for $(i-1)\Delta t \leq \tau \leq i\Delta t$. And $D(\tau) = 0$ for $\tau > I\Delta t$.

The model output will be recorded differently, using the met mobility demand at the end of jth time interval M_j as the output for the jth time interval:

$$M_j = M(j\Delta t) \tag{8}$$

The unit pulse response at $t = j\Delta t$ from an input of duration Δt ending at $(i-1)\Delta t$ is found by equation (6):

$$p[t-(i-1)\Delta t] = p[(j-i+1)\Delta t] = \frac{1}{\Delta t}\int_{(j-i)\Delta t}^{(j-i+1)\Delta t} u(l)dl \tag{9}$$

For $t \geq I\Delta t$, convolution integral equation (4) can be broken down into I parts:

$$M_j = \int_0^{j\Delta t} D(\tau)u(j\Delta t - \tau)d\tau = \sum_{i=1}^{I}\frac{A_i}{\Delta t}\int_{(i-1)\Delta t}^{i\Delta t} u(j\Delta t - \tau)d\tau \tag{10}$$

In each of the I integrals, there is $l = t - \tau = j\Delta t - \tau$, together with equation (9), for the jth integral there is:

$$\begin{aligned}\frac{A_i}{\Delta t}\int_{(i-1)\Delta t}^{i\Delta t} u(j\Delta t - \tau)d\tau &= \frac{A_i}{\Delta t}\int_{(j-i+1)\Delta t}^{(j-i)\Delta t} -u(l)dl \\ &= \frac{A_i}{\Delta t}\int_{(j-i)\Delta t}^{(j-i+1)\Delta t} u(l)dl = A_i p[(j-i+1)\Delta t]\end{aligned} \tag{11}$$

Let $U_{j-i+1} = p[(j-i+1)\Delta t]$, equation (11) can become:

$$M_j = \sum_{i=1}^{I} A_i U_{j-i+1} \tag{12}$$

Similarly, for $t < I\Delta t$, the output can be divided into j parts at time $t = j\Delta t$ written as:

$$M_j = \sum_{i=1}^{j} A_i U_{j-i+1} \tag{13}$$

Combining equation (12) and equation (13), there is:

$$M_j = \sum_{i=1}^{min[I,j]} A_i U_{j-i+1} \tag{14}$$

Equation (14) can be further expressed in matrix form:

$$[A][U] = [M] \tag{15}$$

Linear regression can be used to derive U for equation (15) given A and M. Assume an estimate will be found for U that yields $[\hat{M}]$. The solution can be found with least square error minimization between $[M]$ and $[\hat{M}]$. The solution will be:

$$[U] = [[A]^T[A]]^{-1}[A]^T[M] \tag{16}$$

Equation (1) can be rewritten in discrete time as:

$$S_j - S_{j-1} = A_j - M_j \tag{17}$$

Where S_j is the unmet mobility demand at the end of the jth time interval. And $A_j = 0$ for $j > I$.

Assume the initial unmet mobility demand is zero ($S_0 = 0$), iteratively, equation (17) can become:

$$S_j = \sum_{j=1}^{min[I,j]} A_j - \sum_{j=1}^{j} M_j \tag{18}$$

4 Discussion

To use this SLR model to estimate the unmet mobility demand, three types of data will be needed for a given region: (i). Historical data of total human mobility demand from multiple sources as highlighted (A_{hist}); (ii). Historical data of met mobility demand (M_{hist}); (iii). Future prediction data of total human mobility demand (A_{future}). Use A_{hist} and M_{hist} the 'unit hydrograph' of mobility demand (U) in this region can be found using least-squares fitting with equation (16). The obtained U will be used with A_{hist} for equation (14) to estimate the future met mobility demand M_{future}. Finally, with equation (18) the future unmet mobility demand S_{future} can be estimated.

We provide a recommended work-flow to use the 'unit hydrograph' approach in unmet mobility demand prediction (also see Algorithm 1):

1. Delineate the studied city into sub-areas, so that within each sub-area the total travelling demand is roughly uniformly distributed;
2. For each of the sub-areas, collect multi-source data of total mobility demand and demand that has been met based on relevant historical data-source and/or results from other models;

3. Within each sub-region, develop the 'unit hydrograph';
4. Use the 'unit hydrograh' with the predicted/estimated future mobility demand data to compute the future mobility demand that will be met;
5. Use the computed future mobility demand that will be met together with the future total mobility demand, the future unmet mobility demand can be estimated.

Algorithm 1: Linear Reservoir Model for Unmet Mobility Demand Estimation

input : The historical urban human mobility demand A_i with each time intervals Δt for a particular region, the met urban human mobility demand M_j with the same time and spatial frame as A_i from multi-source historical data, including taxi pick-ups, public transportation usage, bike usage, ect.

output: The future unmet mobility demand S_j

1 Use deconvolution and linear regression to develop the 'unit hydrograph' U with historical A and M data;

2 Use derived U from step 1 to predict future M_j by equation (15), given that future A known (Future A can be estimated from arbitrary transportation demand models);

3 Use A_i and M_j to compute the predicted unmet mobility demand S_j iteratively for each time interval.

Since this is still an ongoing work, we did not implement and compare the SLR model with other unmet demand estimation algorithms [1, 2]. In future work, we will conduct an experiment in NYC, using multi-source data-sets to validate the accuracy of the model as well as calibrate it (see Fig. 2). We believe such a model will provide us more insights in understanding the urban human mobility demand problem, and provide decision support for the design of urban transportation system.

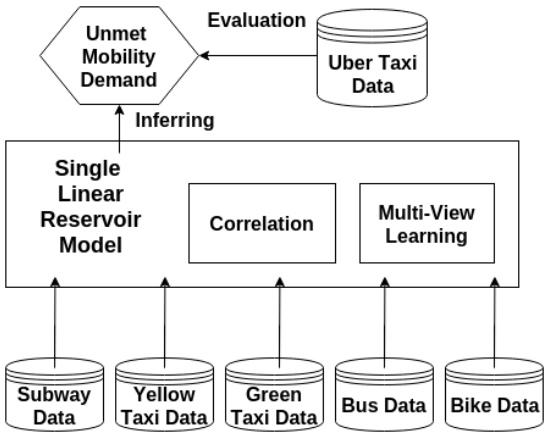


Fig. 2: Framework for Inferring Unmet Mobility Demand.

5 Related Work

5.1 Mobility Modeling based on Multi-Source Urban Data

Existing urban human mobility are mostly driven by data from a single view, e.g., data from a single transportation view such as taxi or subway. The study based on the single-source data inevitably introduces a bias against city residents not contributing this type of data, e.g., residents who walk [8] or ride private vehicles. To address this issue, Zhang et al. [9] propose a human mobility model based on multi-source urban data. They introduce a multi-view learning framework and observe that the model outperforms a single-view model by 51% on average. Zhao et al. [11] decompose the human mobility trips into different classes according to different transportation modes, such as Walk/Run, Bike, Train/Subway or Car/Taxi/Bus. They observe that human mobility can be modelled as a mixture of different transportation modes, and these single transportation movement patterns can be approximated by a log-normal distribution.

5.2 Inferring Unmet Taxi Demand

Recent papers try to infer the unmet taxi demand, the number of people who need a taxi but could not find one, from the taxi data-set. In [2] the authors combine flight arrival with taxi demand and predict the passenger demand at different airport terminals in Singapore use queuing theory. Anwar et. al [1] formalize the unmet taxi demand problem and present a novel heuristic algorithm to estimate it without any additional information. They infer the unmet taxi demand from taxis with empty services and show that it can be used to quantify the unmet demand.

6 Conclusion and Future Work

In this paper we examine the problem of predicting unmet mobility demand with a hydrological perspective. A SLR model is developed for modeling unmet mobility demand. We establish the mathematical relationship between the urban human mobility demand and the unmet mobility demand, and hence develop a 'unit hydrograph' methodology for predicting future unmet mobility demand. In the next step, we will conduct an experiment in NYC, using multi-source data-sets to validate the accuracy of the model as well as calibrate it. We believe such a model will provide us more insights in understanding the urban human mobility demand problem, and provide decision support for the design of urban transportation system.

References

1. A. Afian, A. Odoni, and D. Rus. Inferring unmet demand from taxi probe data. In *2015 IEEE 18th International Conference on Intelligent Transportation Systems*, pages 861–868, Sept 2015.

2. A. Anwar, M. Volkov, and D. Rus. Changinow: A mobile application for efficient taxi allocation at airports. In *2013 IEEE 16th International Conference on Intelligent Transportation Systems*, pages 694–701, Oct 2013.
3. V. T. Chow, D. R. Maidment, and L. W. Mays. *Applied hydrology*. 1988.
4. J. Freire, A. Bessa, F. Chirigati, H. T. Vo, and K. Zhao. Exploring what not to clean in urban data: A study using new york city taxi trips. *IEEE Data Eng. Bull.*, 39(2):63–77, 2016.
5. F. Miranda, H. Doraiswamy, M. Lage, K. Zhao, B. Gonçalves, L. Wilson, M. Hsieh, and C. T. Silva. Urban pulse: Capturing the rhythm of cities. *IEEE Trans. Vis. Comput. Graph.*, 23(1):791–800, 2017.
6. New York open data set. `http://www1.nyc.gov/site/planning/data-maps/open-data.page`.
7. New York Taxi data set. `http://www.nyc.gov/html/tlc`.
8. W. Rao, K. Zhao, Y. Zhans, P. Hui, and S. Tarkoma. Maximizing timely content advertising in dtns. In *9th Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, SECON 2012, Seoul, Korea (South), June 18-21, 2012*, pages 254–262, 2012.
9. D. Zhang, J. Zhao, F. Zhang, and T. He. comobile: real-time human mobility modeling at urban scale using multi-view learning. In *in SIGSPATIAL, Bellevue, WA, USA, November 3-6*, pages 40:1–40:10, 2015.
10. K. Zhao, D. Khryashchev, J. Freire, C. T. Silva, and H. T. Vo. Predicting taxi demand at high spatial resolution: Approaching the limit of predictability. In *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*, pages 833–842, 2016.
11. K. Zhao, M. Musolesi, P. Hui, W. Rao, and S. Tarkoma. Explaining the power-law distribution of human mobility through transportation modality decomposition. *Nature Scientific Reports*, 5(9136), March 2015.
12. K. Zhao, S. Tarkoma, S. Liu, and H. T. Vo. Urban human mobility data mining: An overview. In *2016 IEEE International Conference on Big Data, BigData 2016, Washington DC, USA, December 5-8, 2016*, pages 1911–1920, 2016.