Interpreting Traffic Dynamics using Ubiquitous Urban Data

Fei Wu Hongjian Wang Zhenhui Li
College of Information Sciences and Technology
Pennsylvania State University, University Park, PA, USA
{fxw133, hxw186, jessieli}@ist.psu.edu

ABSTRACT

Given a large collection of urban datasets, how can we find their hidden correlations? For example, New York City (NYC) provides open access to taxi data from year 2012 to 2015 with about half million taxi trips generated per day. In the meantime, we have a rich set of urban data in NYC including points-of-interest (POIs), geo-tagged tweets, weather, vehicle collisions, etc. Is it possible that these ubiquitous datasets can be used to explain the city traffic? Understanding the hidden correlation between external data and traffic data would allow us to answer many important questions in urban computing such as: If we observe a high traffic volume at Madison Square Garden (MSG) in NYC, is it because of the regular peak hour or a big event being held at MSG? If a disaster weather such as a hurricane or a snow storm hits the city, how would the traffic be affected?

Most of existing studies on traffic dynamics focus only on traffic data itself and do not seek for external datasets to explain traffic. In this paper, we present our results in attempts to understand taxi traffic dynamics in NYC from multiple external data sources. We use four real-world ubiquitous urban datasets, including POIs, weather, geo-tagged tweets, and collision records. To address the heterogeneity of ubiquitous urban data, we present carefully-designed feature representations for these datasets. Our analysis suggests that POIs can well describe the regular traffic patterns. In addition, geo-tagged tweets can be used to explain irregular traffic caused by big events, and weather may account for abnormal traffic drops.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; H.4.0 [**Information Systems Applications**]: General

Keywords

Urban computing, traffic

1. INTRODUCTION

Traffic is the pulse of the city that impacts the daily life of millions of people. Traffic congestion can make drivers frustrated and also generate a lot of city noises and vehicle accidents. Therefore, there has been a longstanding strong demand to understand and forecast traffic under different scenarios. An insightful analysis on traffic dynamics could lead to intelligent transportation systems that make the city flow more smooth and make people's life easier.

Modeling traffic dynamics, however, is very difficult as the traffic varies significantly over space and time and it is impacted by many factors simultaneously. To date, various approaches have been proposed to model and to predict traffic [13]. But most of these studies focus on how to predict traffic using *historical traffic data*. For example, the traffic volume at 5 p.m. today will be high because the traffic volume has always been high at 5 p.m. on weekdays; the traffic will increase at a location because nearby locations are experiencing an increasing trend of traffic. Unfortunately, these patterns fail to provide a *semantic understanding* of the traffic. For many intelligent transportation applications, an ideal interpretation typically involves *contextual urban information* and may be of the following form: *Traffic at location A is dominated by the daily routine commute at this place, location B has a significant traffic increase when there is a local concert, and the sudden drop of traffic at these locations on certain day is due to a heavy snow.* Instead of relying on the knowledge of local experts, we argue that such an interpretation could be automatically learnt from the data.

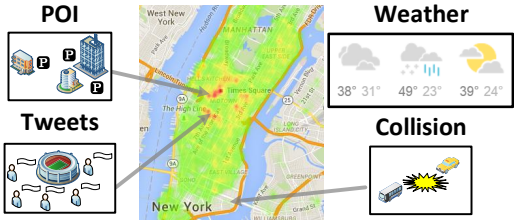


Figure 1: Using ubiquitous urban data to explain traffic.

Motivated by this goal, we propose to study a novel problem: interpreting traffic data using external contextual urban data. Moreover, the big data era has brought us unprecedented urban data, which enables us to take a systematic approach to address this problem. Take New York City for example. The city generates about half million medallion taxi trips; all these data from year 2009 to year 2015 are publicly available on www.nyc.gov under the Freedom of Information Law (FOIL). The NYC taxi data was first made public in 2014. It is the first massive public traffic dataset which contains extremely rich information about the urban dynamics in NYC.

In the meantime, to understand such a large-scale traffic data, several contextual urban data in NYC are being collected from different sources. For example, information about more than 380K

SIGSPATIAL'16, October 31-November 03, 2016, Burlingame, CA, USA

DOI: http://dx.doi.org/10.1145/2996913.2996962

points-of-interest (POI) can be collected from FourSquare; people generate about half million geotagged tweets per day in the city; National Centers for Environmental Information provides daily climate information with 28 weather attributes collected from a monitoring tower in Central Park; more than 769K vehicle collisions from 2012 to today are available on NYC open data website (data.cityofnewyork.us). Our key insight is that all these ubiquitous urban datasets could potentially be valuable signals to explain the traffic dynamics: POIs describe the functions of a region (e.g., a business district typically attracts a large volume of morning taxi drop-offs and evening taxi pick-ups; an area with many nightclubs has increased taxi drop-offs at night and pick-ups after midnight). Geo-tagged tweets capture local events (e.g., a popular event will generate peaks in drop-offs before the event and pick-ups after the event). Extreme weather could lead to traffic decline. Vehicle collisions might cause temporary road closures and traffic jams.

Our paper has three major contributions:

- We study a novel and important problem in urban computing: understanding traffic using ubiquitous urban data.
- We investigate how to design features and models to capture the correlations between traffic and different types of urban data.
- Our experiments show that external datasets can be helpful in interpreting taxi traffic.

The rest of the paper is organized as follow. We review the literature in Section 2. Section 3 describes datasets and how we design features based on the properties of these datasets. Section 4 presents our model. We show the empirical results in Section 5 and conclude our study in Section 6.

2. RELATED WORK

Traffic Prediction. Traffic prediction has been extensively studied in transportation research area. Representative forecasting models include such as neural network models [2], autoregressive integrated moving average (ARIMA) models [1], Bayesian network approach [11], and Markov Random Fields [3]. Instead of forecasting future traffic based on historical traffic data, we aim to model the correlation between traffic and external urban datasets. Our work takes into consideration several large-scale urban datasets such as POI, geo-tagged tweets, weather, and vehicle collisions and study their impacts on traffic.

Computing with Heterogeneous Urban Data. In recent years, urban computing [13] has gain an increasing popularity due to the availability of large-scale urban data. These studies include using different urban datasets such as POI, taxi, bike rental, or noise complaint to profile city functions [12], detect traffic anomalies [16], predict air quality [14, 15], gas consumption [8], location recommendation [5], and use profiling [10, 9]. Our work is under the same theme of urban computing in the context of large-scale heterogeneous data. To the best of our knowledge, there is no prior work on modeling the traffic data using multiple large-scale contextual urban datasets.

3. DATA AND FEATURE DESIGN

Taxi Data. A large-scale New York City taxi dataset has been made public online (http://www.nyc.gov/html/tlc/html/about/trip_record_data.shtml). The dataset contains all trips completed in yellow and green taxis from 2009/01/01 to 2015/12/31. Each trip contains information about pick-up location and time, drop-off location and time, trip distance, fare amount, etc. We use the subset of trips from 2012/10/1 to 2012/12/31, which has **28,759,878** trips. On average, there are 463,869 trips per day. We pick this time period because it aligns with the date of collection of other external context datasets. Due to the space limit, we only report the results on taxi drop-offs in this paper.

Feature: Point-of-Interest (POI) for Regional Functions. We collect a POI dataset from FourSquare API [4]. The FourSquare API provides us with venue information such as venue name, category, number of check-ins, and number of unique visitors. We use the categorical distribution of POI to characterize the neighborhood functions. There are 10 first-level categories in total, such as food, residence, and travel. We follow the same querying strategy as described in [9]. In total, we obtain information about 380,380 venues for New York City.

It is important to consider the popularity over time for POIs. The time-varying popularity indicates the time span during which a region may be of interest to people. We obtain dataset of FourSquare check-in posts from Twitter following the strategy introduced in [9]. As a result, we obtain 1,598,617 check-ins. Later, the check-ins are aggregated by the POI category and by the hour of the day.

Formally, let $\mathbb{C}$, $\mathbb{D}$, and $\mathbb{T}$ denote the set of all categories, the set of all spatial grid cells, and the set of all timestamps, respectively. We consider a set of n POIs on the map: $\mathbb{P} = \{p_1, p_2, \ldots, p_n\}$. Each POI p_i is represented as a tuple (c, l, z), where $p_i.c$ is the category of this POI, $p_i.l$ is its geographic location, and $p_i.z$ is the overall popularity of this POI measured by the total check-ins from FourSquare (this data is directly obtained from FourSquare API). We calculate the POI feature value for category $c \in \mathbb{C}$ in grid cell $d \in \mathbb{D}$ at time $t \in \mathbb{T}$ as:

$$\mathbf{f}_{POI}(c,d,t) = \sum_{i:p_i.l \in d \wedge p_i.c = c} p_i.z \times g(c, P(t)), \tag{1}$$

where $P(t)$ is the relative time (i.e., hour of the day) of t, $g(c,P(t))$ is the temporal popularity of the category c at relative time of t.

Feature: Geo-Tagged Tweets for Local Events. The POI features may help us capture the regular traffic patterns at locations. To further capture irregularity in traffic, we seek to extract event occurrences using geo-tagged tweets. Again we use the geo-tagged tweets we collected in NYC around the same time period (from Oct. 2012 to Dec. 2012). Each geo-tagged tweet is of the form of

$$\langle timestamp, userid, latitude, longitude, content \rangle.$$

Formally, we define the tweet feature value for a grid cell $d \in \mathbb{D}$ at time $t \in \mathbb{T}$ as:

$$\mathbf{f}_{tweet}(d,t) = c(d,t), \tag{2}$$

where $c(d,t)$ is the count of distinct users post a tweet at grid cell d at time t. We count the number of users instead of the posts to alleviate the problem of spammers.

Feature: Weather for Disasters. Intuitively, the traffic should correlate with weather. Furthermore, extreme weather conditions such as hurricane and snow storm could significantly impact traffic. To capture the impact of weather, we use the daily weather dataset in USA from National Centers for Environmental Information (http://www.ncdc.noaa.gov/). Among 28 weather attributes, we use the highest 2-mins wind speed, highest 5-seconds wind speed, precipitation, and snow fall information.

Formally, we define the impact of an extreme weather event e at time t (after the event) on its corresponding weather feature a as:

$$\mathbf{f}_e(a,t) = max\{c(a,t_e) - \lambda(t-t_e)^{\alpha}, 0\}, t \in [t_e, \infty), \tag{3}$$

where α and λ are positive values controlling the decay, t_e is the time when event e happens, and $c(a,t_e)$ is the value of weather feature (corresponding to the extreme weather event) at time t_e. We

let $\alpha > 1$ to capture the lasting effect of severe weather conditions, and set $\mathbf{f}_e(a,t) = 0$ for $t < t_e$. Finally, we can define weather features at time t as: $\mathbf{f}_{weather}(a,t) = \sum_{e \in E_a} \mathbf{f}_e(a,t)$, where E_a is set of all extreme weather events related to feature a.

Feature: Vehicle Collisions for Traffic Jam. Vehicle collisions could potentially cause traffic controls on certain blocks and thus impact local traffic. Details of Motor Vehicle Collisions in New York City provided by the Police Department (NYPD) and can be accessed from NYC Open Data (https://data.cityofnewyork.us/). This data is constantly being updated. We use data from the same period, i.e., from Oct. 2012 to Dec. 2012.

Let $\mathbb{R}$ denote the set of all collisions in our study. Each collision record $r_i \in \mathbb{R}$ is represented as a tuple (t,l,s), where $r_i.t$ the time of the collision, $r_i.l$ is the latitude and longitude of the collision, and $r_i.s$ is the number of people injured or killed in this collision. We construct our feature for collisions on day t in grid cell d using the number of collisions weighted by the severity of the collisions:

$$\mathbf{f}_{collision}(d,t) = \sum_{i:r_i.t=t \wedge r_i.l \in d} (r_i.s+1). \quad (4)$$

4. MODEL

In this section, we model the correlation between traffic and urban features. For each location grid d, we denote the number of pick-ups at time t as y_t. The feature space is a combination of all the features:

$$\mathbf{x}_t = [\mathbf{f}_{POI}(:,d,t), \mathbf{f}_{tweet}(d,t), \mathbf{f}_{weather}(:,t), \mathbf{f}_{collision}(d,t)].$$

Here, $\mathbf{f}_{POI}$ is a 10-dimensional feature because there are 10 categories for the POIs, $\mathbf{f}_{tweet}$ and $\mathbf{f}_{collision}$ are one-dimensional features, and $\mathbf{f}_{weather}$ is k-dimensional feature where k depends on the number of weather attributes used. In our experiment setting, we use wind speed, precipitation, and snow fall. Given the data $\{y_t, \mathbf{x}_t\}_{t=1}^N$ for a location grid d, our goal is to fit a regression model $y = f(\mathbf{x})$ for that grid.

Linear regression is the most frequently used technique to fit the data. But it could suffer from the overfitting issue. A common technique to avoid overfitting is to control the values of weights $\mathbf{w}$ on the features, where we add an L_2-regularization to the objective function:

$$J(\mathbf{w}) = (\mathbf{y} - X\mathbf{w})^T(\mathbf{y} - X\mathbf{w}) + \lambda \|\mathbf{w}\|^2, \quad (5)$$

where X is the $N \times D$ design matrix. In addition, the features may have non-linear (e.g., multiplicative) interactions among them. For example, in a bad weather day, the overall traffic volume would decrease, but the relative traffic pattern over time (e.g., peak hours vs. non-peak hours) remains similar, which can be described by the temporal popularity of POIs. In this case, the traffic is determined as a combined effect of the weather feature and the POI feature. Therefore, we further adopt the kernel ridge regression (KRR) model [7] with a degree-2 polynomial kernel, which consider the interactions between features as a multiplication of any two features.

5. EMPIRICAL RESULTS

5.1 Quantitative Study

We divide Manhattan into 500 meter $\times$ 500 meter grids and build a model for each grid to predict the hourly number of taxi drop-offs. There are 319 grids in the middle and lower area of Manhattan (i.e., South of 86th street). Our experiments are conducted on these grids where the data from different sources are less sparse. We use the first two weeks as training (from 10/01/2012 to 10/15/2012) to fit the model and use the following week for testing (from 10/16/2012 to 10/23/2012). The average hourly drop-off frequency is 53 for all grids and the standard deviation for hourly drop-off is 38. The traffic also varies for different grids with the highest hourly drop-off being 926 and lowest hourly drop-off being 0. We use mean-square error (MSE) and mean relative error (MRE) to evaluate testing error (on testing data) and coefficient of determination (R^2) to evaluate the fitness of the model (on training data). To study feature importance, we use leave-one-out strategy, that is, testing the model performance by excluding one feature from the feature set. Kernel ridge regression is the model used in this experiment.

	all	T,W,C	P,T,C	P,W,C	P,W,T
R^2	0.64	0.07	0.63	0.62	0.61
RMSE	24.1	50.9	24.1	23.7	24.1
MRE	0.5	1.51	0.5	0.5	0.5

Table 1: Feature effectiveness. P: POIs, T: geo-tagged tweets, W: weather, C: collision. RMSE, MRE and R^2 values are reported for different combination of features.

Table 1 summaries the model performance with different features on the taxi drop-off data. Without using POI features, RMSE is nearly two times larger than the model with POI features included, and MRE is three times larger. Similarly, R^2 is 0.06 when excluding POI features, while R^2 is always higher than 0.6 when POI features are used. These results suggest that POI features correlate the best with the taxi traffic data in general. At the same time, there is no significant improvement by including weather data, collision data, and tweet data. This is because traffic follows routine behavior for most of the time, which is captured by the POI features. Other features can describe abnormal scenarios (e.g., events and extreme weather), but such scenarios happen less frequently or only happen in some small regions (e.g., a convention center). In the following section, we qualitatively investigate several regions of interest to understand the correlation between traffic and these features.

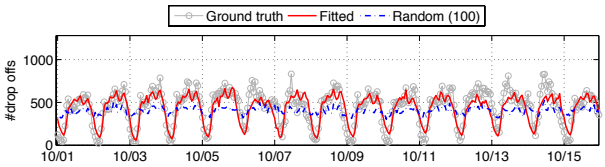
5.2 Qualitative Analysis

We select four regions for qualitative analysis, i.e., Times Square (a tourist sight with entertainment related venues), Madison Square Garden (a multi-purpose arena; a transportation center Penn Station sits beneath MSG), 5th avenue (a central business area with many shops), and Jacob K. Javits Convention Center (a large convention center). We design our analysis with two questions in mind:

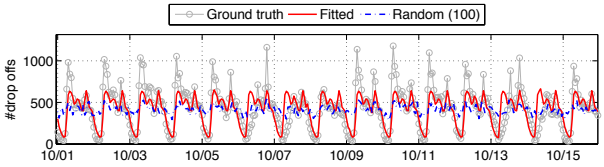
- Q1 (Fitness): Can we construct the taxi traffic patterns by using external urban data?
- Q2 (Interpretation): Which features are useful and under what circumstances they are useful?

To answer these two questions, we follow a similar methodology carried out in previous study [6] and look at the fitting results of our model on all areas.

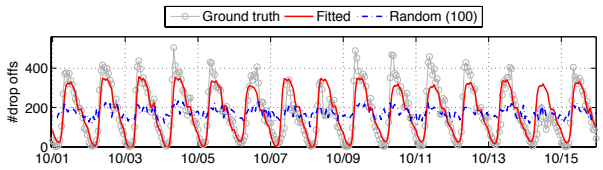
Figure 2(a), (b), and (c) show the fitting results using our POI features for drop-offs at Madison Square Garden, Times Square, and 5th Avenue, respectively. For comparison, we generate 100-dimensional features with random values and use the same kernel ridge regression model to fit the traffic data (shown as blue dashed line in Figure 2(a), (b), and (c)). Obviously, these random features, although with a much higher dimension, cannot fit the traffic data well. This demonstrates the effectiveness of our POI features in explaining the traffic.



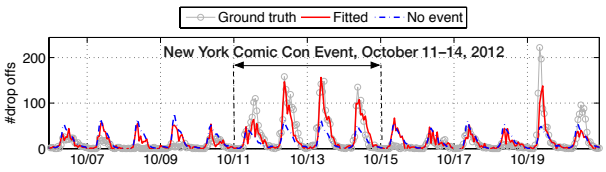
(a) Madison Square Garden. Traffic explained by POI.



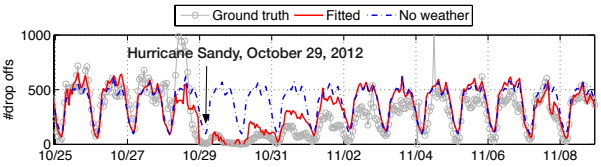
(b) Times Square. Traffic explained by POI.



(c) Fifth Avenue. Traffic explained by POI.



(d) Javits Convention Center. Traffic explained by POI and geo-tagged tweets.



(e) Madison Square Garden. Traffic explained by POI and weather.

Figure 2: Fitting results of traffic data using other urban data.

Geo-tagged tweets may help explain unexpected traffic patterns where local events are the dominant source of traffic. Although overall geo-tagged tweets do not correlate with traffic, we found cases where traffic patterns can be attributed to ongoing events at the venue of large events, i.e., Javits center. Figure 2(d) shows the traffic data at Javits center during the period from 10/11/2012 to 10/13/2012, when there is large event, i.e., NYC Comic Con. It is clear that including geo-tagged tweets significantly improve the fitness of our model for the event days.

Intuitively, the extreme weather condition should impact the traffic. Here we particularly look at one weather event occurred during the time covered by our datasets. Hurricane Sandy is a category-3 major hurricane that hit New York City on Oct. 29, 2012. The wind speed attribute in our weather feature captures the signal of this disaster. As shown in Figure 2(e), incorporating the weather information can effectively capture the significant drop of traffic volume during the time. By modeling the recovery time after the disaster, the fitted traffic pattern shows a slow recovery of traffic in the next 3 days, which aligns with the actual traffic pattern. Compared with using the POI features only, it demonstrates the utility of weather data in the presence of extreme weather conditions.

We also find collision feature does not show any significant correlation with traffic. The reason could be that, wile the collisions do affect the local traffic, its impact is subtle on the overall traffic (in terms of number of pick-ups and drop-offs). Capturing such minor impact using urban data still remains a challenging problem.

6. CONCLUSION

In this paper, we explore the potentials of using ubiquitous urban datasets to interpret traffic data. We use a large-scale taxi trip data in New York City. The explanatory urban datasets include POIs, geo-tagged tweets, weather, and vehicle collisions. We propose to use kernel ridge regression to describe the non-linear non-additive relationships of impacting factors. We demonstrate that using ubiquitous urban datasets can help us better understand the urban dynamics, which could potentially benefit a set of applications such as smart city and intelligent transportation system.

Acknowledgements

The work was supported in part by NSF award #1618448 and award #1544455. The views and conclusions contained in this paper are those of the authors and should not be interpreted as representing any funding agencies.

7. REFERENCES

[1] A. Abadi, T. Rajabioun, and P. A. Ioannou. Traffic flow prediction for road transportation networks with limited traffic data. *Intelligent Transportation Systems, IEEE Transactions on*, 16(2):653–662, 2015.

[2] K. Y. Chan, T. S. Dillon, J. Singh, and E. Chang. Neural-network-based models for short-term traffic flow forecasting using a hybrid exponential smoothing and levenberg–marquardt algorithm. *Intelligent Transportation Systems, IEEE Transactions on*, 13(2):644–654, 2012.

[3] P.-T. Chen, F. Chen, and Z. Qian. Road traffic congestion monitoring in social media with hinge-loss markov random fields. In *ICDM'14*. IEEE, 2014.

[4] Fourquare. https://foursquare.com/, 2016.

[5] D. Lian, Y. Ge, F. Zhang, N. J. Yuan, X. Xie, T. Zhou, and Y. Rui. Content-aware collaborative filtering for location recommendation based on human mobility data. In *ICDM'15*. IEEE, 2015.

[6] Y. Matsubara, Y. Sakurai, W. G. Van Panhuis, and C. Faloutsos. Funnel: automatic mining of spatially coevolving epidemics. In *KDD'14*. ACM, 2014.

[7] K. P. Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[8] J. Shang, Y. Zheng, W. Tong, E. Chang, and Y. Yu. Inferring gas consumption and pollution emission of vehicles throughout a city. In *KDD'14*. ACM, 2014.

[9] F. Wu and Z. Li. Where did you go: Personalized annotation of mobility records. In *CIKM'16*. ACM, 2016.

[10] F. Wu, Z. Li, W.-C. Lee, H. Wang, and Z. Huang. Semantic annotation of mobility data using social media. In *WWW'15*, 2015.

[11] Y. Xu, Q.-J. Kong, R. Klette, and Y. Liu. Accurate and interpretable bayesian mars for traffic flow prediction. *Intelligent Transportation Systems, IEEE Transactions on*, 15(6):2457–2469, 2014.

[12] J. Yuan, Y. Zheng, and X. Xie. Discovering regions of different functions in a city using human mobility and pois. In *KDD'12*, 2012.

[13] Y. Zheng, L. Capra, O. Wolfson, and H. Yang. Urban computing: concepts, methodologies, and applications. *TIST*, 5(3):38, 2014.

[14] Y. Zheng, F. Liu, and H.-P. Hsieh. U-air: When urban air quality inference meets big data. In *KDD'13*. ACM, 2013.

[15] Y. Zheng, X. Yi, M. Li, R. Li, Z. Shan, E. Chang, and T. Li. Forecasting fine-grained air quality based on big data. In *KDD'15*. ACM, 2015.

[16] Y. Zheng, H. Zhang, and Y. Yu. Detecting collective anomalies from multiple spatio-temporal datasets across different domains. In *GIS'15*. ACM, 2015.