Proc. of the 6th Intl. Symposium on Reliability Engineering and Risk Management (6ISRERM)
31 May – 1 June 2018, Singapore
Editor(s) Xudong Qian, Sze Dai Pang, Ghim Ping Raymond Ong, Kok-Kwang Phoon

ANALYSIS OF TRANSPORT SYSTEM FLOWS USING VAST MOBILE PHONE DATA

Shaoweihua Liu, Anish Khadka, Yang Liu, Zhiyuan Liu*

Jiangsu Key Laboratory of Urban ITS, Jiangsu Province Collaborative Innovation Center of Modern Urban Traffic Technologies, School of Transportation, Southeast University, China* corresponding author, Email: zhiyuanl@seu.edu.cn

Abstract

Forecasting network flows accurately is of great importance to transportation management. To overcome the shortcoming of a single data source that cannot consider the accuracy and breadth simultaneously, multi-source transportation data, including the mobile phone data, license plate recognition (LPR) data, etc., are used to forecast the traffic flows with big data analytics. Mobile phone data are associated with wide coverage as well as low acquisition costs and are usually used in OD estimation between large zones, commuting travel characteristics analysis and highway flow prediction. However, due to the low precision of the data, its utilization in analyzing the traffic flows in urban roads with high density is difficult. In contrast, LPR data are high precision data, but with relatively low coverage. In this paper, we incorporate these two data sources and utilize their advantages comprehensively based on machine learning approaches to forecast traffic flows in urban roads. More specifically, the LPR data are taken as supervision label while the input features are extracted from mobile phone data with full consideration of spatial-temporal characteristics. Random forests algorithm is developed to forecast network flows with a 90% accuracy.

Keywords: Network flow forecasting; Mobile phone data; Multi-source data fusion; Random forests

1. Introduction

The development and deployment of Intelligent Transportation System (ITS) have significantly revolutionized the various aspects of traffic management through its primary functional areas like Advanced Traffic Management System (ATMS), Advanced Traveler Information System (ATIS), etc., thus enabling travelers to make informed travel decisions while alleviating the externalities of the transportation sector. However, the success and effectiveness of these strategies depend on the availability of real-time information regarding network-wide operation and the constantly evolving traffic conditions, which is directly reflected by traffic flows. The past research mainly concentrated on responsive schemes responding to previous traffic conditions, neglecting the importance and need for real-time adjustment of management strategies (Habtemichael et al., 2016). Short-term traffic flow forecasting methods are proposed to help the decision makers develop an intuitive understanding of future traffic status and make timely adjustments accordingly.

The advances in sensor technology and its wide applications in ITS have generated a large amount of transportation data, which can be directly utilized in forecasting short-term traffic flow (Li et al., 2015). The sensing units can be divided into the following two types namely; fixed sensing units and mobile sensing units. The former is more prevalent due to the instantaneity and accuracy in information acquisition but requires the occupation of road facilities and expensive to install and maintain, thus limiting its large-scale application in the urban networks (Gao et al., 2013). Among them, License Plate Recognition (LPR) data provide high definition information of vehicles and avoid low penetration rate that generally occurs in other sensing units. The latter type can be subdivided into Global Positioning System (GPS) based sensing units and mobile phone based sensing units, which have successfully compensated the limitations of fixed sensing units. GPS based sensing units have significantly small positioning error and relatively stable positioning time interval, however, the lack of sufficient number of GPS equipped vehicles results in potentially biased actual traffic conditions. The advancement

in mobile technology has made the information of personal mobility patterns easier to access while the data collected from the mobile phone are associated with greater scale and coverage (Lane et al., 2010). Unfortunately, the deviation of mobile phone position varies from 150m to 500m, making it difficult for network flow analysis in urban roads with high density.

With the objective of compensating defects of individual data source, multi-sensor data fusion (DF) techniques are proposed for combining multi-sensors to regenerate a dataset (Kessler, 1992), enhancing the confidence, robustness and spatial coverage of the datasets. Therefore, the same result derived from multi-source data can authenticate mutually to reduce the uncertainty (El Faouzi et al., 2011). Additionally, the expansion of spatial and temporal range can be achieved through DF process with different coverage rate (Zhu et al., 2016). Present DF methods applied in the transportation system can be divided into three main categories namely; statistics, probability and artificial intelligence (El Faouzi et al., 2011; Han et al., 2000; Dubois et al., 1988). Ensemble learning algorithm, as one of the artificial intelligence, has been proved to have better generalization ability in both classification and regression problems.

In the working paper *Transport Network Flow Estimation Based on the Multi-Source Big Data: A Supervised Machine Learning Approach*, multiple algorithms, including random forests algorithm (RF) and gradient boosting decision tree algorithm (GBDT), are first proposed to forecast traffic flows in urban areas, and spatial-temporal features are augmented by sliding window (SW) method and predicting accuracy is improved with multi- grained (MG) features. This paper mainly discusses the data preprocessing procedure and predicting performance of random forests algorithm. More specifically, the LPR data are taken as supervision label while the input features are extracted from mobile phone data considering both spatial and temporal characteristics. Random forests models are developed for roads by applying both LPR data and mobile phone data. The traffic flows in the remaining roads with only mobile phone data available are then predicted by adopting model considering similar input features and road structure. The contributions of this paper are twofold. Firstly, to the best of our knowledge, no studies have been conducted analyzing urban transportation network flows using mobile phone data, for which this study makes up the vacancy in the application field of mobile phone data. Secondly, the forecasting results revealed that random forests algorithm has good performance in forecasting traffic flows in urban areas, with prediction precision higher than 90%.

2. Data Preparation

2.1 *Data description*

The LPR data and the mobile phone data are the original data sources chosen to be fused in a supervised machine learning approach based on random forests algorithm. The dataset contains about 3 million LPR data per day and 60 million anonymized mobile phone data per day in the area within a radius of 5 kilometers around Nanjing South Railway Station, which spans 6 days, from October 17, 2016 to October 22, 2016. The LPR data, provided by Jiangsu Information Center, China, have at least 95% recognition accuracy in the daytime and 90% in the nighttime (except motorcycle license plate, temporary license plate and tractor license plate) as per the national standards GA/T 497. The mobile phone data, provided by China Telecom, Jiangsu Branch, own about 30% market share.

Table 1: Samples of LPR data.

Vehicle ID	Access Time	Plate Color	Point No	Lane No	Speed
1261001772	2016/10/19 11:03:17	3	1092	1	32.2

Table 1 depicts the data structure of LPR data. Vehicles can be only tracked through Vehicle ID while the vehicle type is marked by Plate Color. The temporal information can be directly obtained through Access Time while the spatial information needs to be further inferred by Point No and Lane No. The speed is the rough estimation of the spot speed when vehicles approach a checkpoint.

Table 2: Samples of mobile phone data.

IMSI	Access Time	Latitude	Longitude
460110120767844	2016/10/19 17:33:25	31.96077	118.797097

Typical mobile phone data are shown in table 2. To safeguard personal privacy, users' detail information is uniquely identified with IMSI. Access time, recorded as timestamp, indicate the occurrence time of network request of originating

terminals. Coordinates constituted by longitude and latitude describe the location of originating terminals, which are estimated by telecom base station according to standard triangulation algorithm, with an accuracy of about 150-300 m.

2.2 *Data cleansing*

Noise data need to be eliminated to guarantee the cleanliness and completeness of the raw dataset. As a relatively accurate data source, few noise data exist in the LPR data. Contrarily, such data are found in mobile phone data, which can be divided into two main parts namely; false switching data and repeated positioning data. Common false switching data are generated when frequently switching among several adjacent base stations occurs due to the dramatic changes in signal strength, which is well known as ping-pong data. Another false switching data called drift data are generated when the location of mobile phone occasionally jumps to a remote base station and are recorded by adjacent base stations for a short period. Furthermore, numerous repeated positioning data are produced for mobile phones interacting with the same base station for a long time, which should be reduced to an acceptable amount.

Step 1: Trip chain for IMSI i is extracted to form a subset S_i. For consecutive data $(lng_{i,j}, lat_{i,j}, t_{i,j})$, $(lng_{i,j+1}, lat_{i,j+1}, t_{i,j+1})$, straight-line distance and straight-line speed are as follows (Husár et al., 2017):

$$D_{i,j} = 2 \cdot R \cdot \arcsin\sqrt{\sin^2\left(\pi\Delta lat_j/180\right) + \cos(\pi lat_{i,j+1}/180)\cos(\pi lat_{i,j}/180)\sin^2\left(\pi\Delta lng_j/180\right)} \tag{1}$$

$$v_{i,j} = D_{i,j}/\Delta t_{i,j} \tag{2}$$

where, $\Delta lat_j = (lat_{i,j+1} - lat_{i,j})/2$, $\Delta lng_j = (lng_{i,j+1} - lng_{i,j})/2$, $\Delta t_{i,j} = t_{i,j+1} - t_{i,j}$, R is Earth radius (6371 km).

Step 2: Discriminant variables $v'_{i,j}$ and $\Delta t'_{i,j}$ are proposed to distinguish whether noise data occur or not and what kind of abnormal type they belong with, which are defined as:

$$v'_{i,j} = v_{i,j}/3.6 + \theta_v/\Delta t_{i,j} \tag{3}$$

$$\Delta t'_{i,j} = \Delta t_{i,j} + 3.6\theta_t/v_{i,j} \tag{4}$$

where, θ_v and θ_t are penalty coefficients. The threshold v'_T and $\Delta t'_T$ can be calculated as follows:

$$v'_T = v_{\max}/3.6 + \theta_v/\Delta t_{\min} \tag{5}$$

$$\Delta t'_T = \Delta t_{\max} + 3.6\theta_t/v_{\min} \tag{6}$$

where, $v_{\max}$, $v_{\min}$, $\Delta t_{\max}$ and $\Delta t_{\min}$ are the maximum acceptable speed (km/h), minimum acceptable speed (km/h), maximum acceptable duration time (s), and minimum duration time (1s) respectively. θ_v and θ_t can be calibrated by:

$$\theta_v = v_{\max} \cdot \Delta t_{\min}/3.6r \tag{7}$$

$$\theta_t = v_{\min} \cdot \Delta t_{\max}/3.6r \tag{8}$$

where, r is the weight ratio (in this paper, $r = 1$). Data either $v'_{i,j} > v'_T$ or $\Delta t'_{i,j} > \Delta t'_T$ are distinguished as noise data, more specifically, false switching data and repeated positioning data.

2.3 *Data mapping*

Data mapping is necessary for building the relationship between dataset and transportation networks. The spatial information of LPR data is recorded by the location of checkpoints, which is considered as error-free positioning. However, it is difficult to decide whether mobile phone data are correctly located at the actual position or not since the positioning deviation is significantly high, ranging around hundreds of meters. Fortunately, such deviation is uniformly distributed in all data, therefore, the aggregation results are considered to be reliable. The transportation networks should be established before data mapping. Given the coordinates (denote by longitude and latitude) of network key points (extremities or turning points), for a given road segment, the linear interpolation method is used to fill the coordinates between two key points. Assume the coordinates of two key points are (lat_1, lon_1) and (lat_2, lon_2), the coordinates of the ith interpolation point can be obtained as follows;

$$lat_i = k \cdot i \cdot lat_1/D_i + (1 - k \cdot i/D_i) lat_2 \tag{9}$$

$$lng_i = k \cdot i \cdot lng_1/D_i + (1 - k \cdot i/D_i) lng_2 \tag{10}$$

$$D_i = 2 \cdot R \cdot \arcsin\sqrt{\sin^2(\pi \Delta lat/180) + \cos(\pi lat_1/180)\cos(\pi lat_2/180)\sin^2(\pi \Delta lng/180)} \tag{11}$$

where, lat_i and lng_i are the latitude and longitude of the ith interpolation point respectively, k is the interpolation precision (e.g. 50m), $\Delta lat = (lat_2 - lat_1)/2$, $\Delta lng = (lng_2 - lng_1)/2$, R is Earth radius (6371 km).

Mobile phone data are then mapped into the closest point by solving the following optimization problem:

$$\arg\min_x \arcsin\sqrt{\sin^2(\pi(lat_m - lat_x)/360) + \cos(\pi lat_x/180)\cos(\pi lat_m/180)\sin^2(\pi(lng_m - lng_x)/360)} \tag{12}$$

where, lat_x, lng_x, lat_m and lng_m are latitude and longitude of network point and mobile phone data respectively.

2.4 *Data aggregation*

To extract features or to form the supervisory values, data are aggregated in 1-minute time interval by:

$$x_i^{(k)} = \sum_{j=1}^{m} x_{ij}^{(k)} \tag{13}$$

where, $x_i^{(k)}$ is the aggregate number of mobile phone data in road segment i for the kth time interval, $x_{ij}^{(k)}$ is the aggregate number of mobile phone data mapped into the jth point of road segment i for the kth time interval and m is the number of points in road segment i.

Data filling strategy is adopted to ensure the completeness of dataset. For single missing data, the mean value of adjacent two values are used while for multiple missing data, values of the previous day/hour with the same period are used.

2.5 *Feature selection*

The resulting aggregated data contain both spatial and temporal information, which should be taken into full consideration when extracting features. Table 3 shows the various features that are used in this paper.

Table 3: Features extracted from aggregated mobile phone data.

Feature type	Features
Temporal features	Day of week
	Time of day (time slice)
Spatial features	Road segment
Aggregation features	Total number in a time slice

3. Methodology

3.1 *Decision tree*

Decision tree models consider the interpretability while the *if-then rules* emulating human decision process generates set of rules that are easily grasped by non-professional users as well. A decision tree contains one root node, several internal nodes, and leaf nodes. The leaf nodes represent decision results, while internal nodes correspond to several attribute tests. Samples associated with each node are divided into several sub-nodes according to the result of attribute test.

Decision Tree Algorithm:

Input: training data set $D = \{(x_i, y_i)\}_{i=1}^{n}$; attribute set $A = \{a_i\}_{i=1}^{d}$; Function TreeGenerate(D, A).

Algorithm:

if samples in D belong to the same category C **then** classify node as leaf node with label C; **return**

if $A = \emptyset$ **or** samples in D have the same value on A

then classify node as leaf node with the label of the category containing the most samples; **return**

select optimal partition attribute a_* from A;

for every value in a_* **do**

generate a branch node for each node;
define D_v to express the subset of samples values a_*^v in D on a_*;
if D_v is null
then label branch node as leaf node with the label of the category containing the most samples; **return**
else set $\text{TreeGenerate}(D_v, A\backslash\{a_*\})$ as branch node
end

3.2 *Random forests*

Random forests have gained huge popularity due to the good classification and regression performance, robustness, scalability, and ease of use. Multiple decision trees, viewed as weak learners, are ensembled to build a more robust strong learner, random forest, with low generalization error and less susceptible to overfitting (Breiman, 2001).

Random Forest Algorithm:

Input: training data set $D = \{(x_i, y_i)\}_{i=1}^n$ with p input variables;
Algorithm:
initialize: determine M trees to be generated and the number of variables v used for an individual tree ($v < p$);
for $m = 1$ to M **do**
draw a random sample S^* of size n with replacement from D;
loop until (the minimum node size is reached)
for the terminal node of the tree
randomly select v variables out of the p variables;
select the best pair of split variable among the v variables;
split the node into two branch nodes;
output the constructed tree $T_v(x)$;
end

4. Case Study

Based on the aggregated data of area within the radius of 5 kilometers around Nanjing South Railway Station, experiments are conducted to test the performance of the proposed model. The linear interpolation results are shown in figure 1. Original data are aggregated into 8640 samples with 1-min time interval. A total of 864 samples (6 samples per hour * 24 hours * 6 days) are generated for 10-min time interval. However, the sample size is sharply decreased to 288 and 144 for 30-min time interval and 60-min time interval respectively. Therefore, sliding window (SW) method has been adopted to generate more samples. As can be seen in figure 2, assume we need samples with 30-min time interval, the window size is set as 30 and the sliding step is set as 1, then a total 8611 samples (8640-30+1) are generated. For 60-min time interval, the sample size is 8581. The samples are randomly divided into training set and validation set with the ratio of 0.8:0.2.

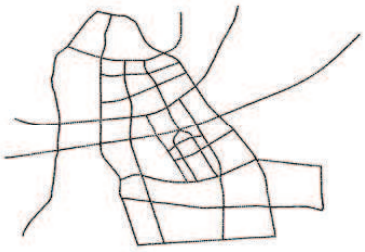

Figure 1: Linear interpolation results.

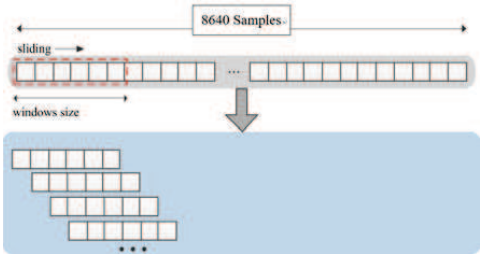

Figure 2: Illustration of sliding window.

To evaluate the performance of proposed method, the Mean Absolute Percentage Error (*MAPE*) is used which is given as:

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{\hat{y}_i} \tag{14}$$

where, y_i is the observed value (aggregation results of LPR data) in the ith time slice, and $\hat{y}_i$ is the estimated value in the ith time slice. N is the number of all estimated values.

In the model training stage, parameters are tuned to achieve the lowest *MAPE*. The learning rate, the maximal number of iterations, maximum depth of individual regression tree and the maximal number of splits are set as 0.01, 320, 5 and 16 respectively. The results are cross-validated in the 5-fold mode by calculating respective *MAPE*. Table 4 shows the performance of models with different time intervals. The result of 10-min time interval achieves the highest forecasting error. The reason is that randomness effect may occur when the traffic flow is particularly low, and a small difference between observed value and forecasting value will lead to a significant relative error. Different model reaches the lowest error with 30-min time interval simultaneously, which shows the existence of optimum time interval. The sliding window method is applied to generate sufficient samples, which can further improve the model performance.

Table 4: Performance of model with different time intervals (*MAPE*).

Model	10-min interval	30-min interval	60-min interval
RF	0.1037	**0.0848**	0.0974
RF +SW	0.1015	**0.0835**	0.0859

5. Conclusion

Forecasting network flows is of vital importance to transportation evaluation and management. Mobile phone data, with wide coverage and low acquisition costs, are of great potential value. However, restricted by the low positioning precision, it is difficult to use mobile phone data to analyze the traffic flows in urban roads with high density. Contrarily, license plate recognition (LPR) data, with high positioning precision, are limited by the relatively low coverage. To overcome these limitations, multi-source data fusing theory has been adopted to take both accuracy and breadth into consideration. In this paper, supervised machine learning approach has been developed to forecast network flows in urban roads by utilizing the advantages of two types of data comprehensively. Spatial-temporal characteristic of mobile phone data is taken into full consideration. Furthermore, a random forests model is proposed, and the aggregation result of mobile phone data and LPR data are considered as features and labels. Sliding window method has been adopted to expand the sample size. The case study of the Nanjing South Railway Station is conducted to validate the performance of the proposed model. The results show that the random forests model has superior performance in forecasting network flows in the urban area, with the best *MAPE* of 10.15%, 8.35% and 8.59% for 10-min, 30-min and 60-min time interval respectively.

References

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5-32.

Dubois, D., & Prade, H. (1988). *Possibility Theory*. Springer US.

Faouzi, N. E. E., Leung, H., & Kurian, A. (2011). *Data fusion in intelligent transportation systems: Progress and challenges - A survey*. Elsevier Science Publishers B. V.

Gao, H., & Liu, F. (2013). Estimating freeway traffic measures from mobile phone location data. *European Journal of Operational Research,229*(1), 252-260.

Habtemichael, F. G., & Cetin, M. (2016). Short-term traffic flow rate forecasting based on identifying similar traffic patterns. *Transportation Research Part C, 66*, 61-78.

Han, J., & Kamber, M. (2001). *Data Mining Concept and Techniques*.

Husár, L., Švaral, P., & Janák, J. (2017). About the geometry of the Earth geodetic reference surfaces. *Journal of Geometry and Physics*, *120*, 192-207.

Kessler, J. (1992). Functional description of the data fusion process. *Japanese Journal of Psychiatric Rehabilitation, 11*.

Lane, N. D., Miluzzo, E., Lu, H., Peebles, D., Choudhury, T., & Campbell, A. T. (2010). A survey of mobile phone sensing. *IEEE Communications Magazine, 48*(9), 140-150.

Zhu, L., Guo, F., Polak, J. W., & Krishnan, R. (2017). Multisensor Fusion Based on Data from Bus GPS, Mobile Phone, and Loop Detectors in Travel Time Estimation. *Transportation Research Board, Meeting*.