Estimating Beijing's Intersection Delays from Floating Car Data: A Boosting Approach

Xiliang LIU, Feng LU, Hengcai ZHANG

State Key Laboratory of Resources and Environmental Information system, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China
{liuxl, luf, zhanghc}@lreis.ac.cn

1. Introduction

Intersections are important components in the urban road networks which demonstrate the spatial-temporal characters of city dynamics and contribute much to the total travel time cost (Nielsen et al., 1998). Intersection delays, which can be defined as the turn cost that is related to the continuation of travel in a node (Winter, 2002), constitute a large part of travel times on urban links (Viti et al., 2004). Nielsen et al. conclude that intersection delays account for 17~35% of the total travel time, according to a survey conducted in the medium congested Municipality of Copenhagen (Nielsen et al., 1998). However, as Zheng et al. have discussed, although intersection delays are the key factor which influences travel time estimation and forecast, most of researches ignore the existence of intersection delays (Zheng et al., 2010).

One possible explanation is that the data from survey or field experiment is hard to obtain and update at the early age. Some researchers rely on theoretical approaches and sophisticated micro-simulation (Horowitz, 1997; Aashtiani et al., 1999; Ding, 2007). These methods might suit for solving intersection delays from a theoretical perspective, but it is easy to draw the wrong conclusions and fall short of the ability to generally imitate the real world if one is not fully familiar with the models (Long et al., 2011). With the development of stationary sensors (inductive loop detectors, ultra-sonic vehicle detectors, and closed circuit television cameras, etc.), some researchers seek to estimate intersection delays directly from the vehicle records (Oh, 2003; Ritchie et al., 2005; Kwong et al., 2009). But it should be noticed that the low density of the traditional fixed-point data collecting equipment make it impossible to produce reliable information about travel time within the network (Günnemann et al., 2004). Recently, real-time floating car data (FCD) collected by operating vehicles (taxicabs, probe cars, buses, private cars, etc.) equipped with GPS-enabled computers has become the mainstream in traffic study because of its cost-effectiveness, wide coverage and flexibility compared with other traffic data sources (Herring et al., 2010). Nevertheless, up to now, there are only a few literatures focusing on intersection delay estimation based on FCD. Some researchers utilize historical mean method to calculate the intersection delays (Sun, 2007; Zhao, et al., 2013). This method demands there exists enough historical FCD records in a given time period so that all the records obey the normal distribution, which seems hard to guarantee. Some researchers employ piecewise linear interpolation to obtain the intersection delays (Ban et al., 2009; Zhang et al., 2011). This solution emphasizes the continuity between consecutive time windows, but has problems when facing to the data sparseness and data missing problems. For the over saturated condition, large calculation error is inevitable

using this method because the variation of the speed is slight. To date, it's hard to get a satisfying estimation result just from one model due to the complexity of transport system. Estimating intersection delays based on FCD regardless of different traffic patterns (peak hour or normal condition) has become a challenging task..

Research in various fields has strongly suggested that the performance of prediction/estimation can be enhanced when (sometimes even in simple fashion) models are combined (Yang et al., 2004). Some linear combining methodologies including equal weights method (Jose et al., 2008), optimal weights method (Granger et al, 1984), minimum error method (Yu et al., 2005), and minimum variance method (Lilian et al., 2000), are applied to determine the weights used in the combined models. However, this kind of methods treats all the inputs at one time, and the calculated weight matrix cannot evolve with new data. The Boosting method, which is a branch of ensemble learning theory in machine learning domain, provides a new way to aggregate different models (Zhou, 2009). Boosting refers to a general and provably effective method of producing a very accurate prediction rule by combining rough and moderately inaccurate rules of thumb in a manner similar to that suggested above. The weight distribution in the Boosting process is not linear combined. It is treated with logarithmic rule which can be derived from every iteration, and evolves along the input data. Empirical observations show that the Boosting method does not suffer from over fitting problem, and has good generalization ability (Opitz et al., 1999).

In this paper, we utilize six different models which have been frequently adopted in previous studies, including linear least squares regression(LLSR), autoregressive moving average (ARMA), historical mean (HM), Artificial Neural Network (ANN), Radical Basis Function Neural Network (RBF-NN), Support Vector Machine (SVM). The first three methods belong to parametric techniques, while the others belong to nonparametric ones. We adopt the Boosting approach to aggregate these six models and compare the estimation result with other four linear combination methods mentioned above. In order to evaluate the model's quality, we first choose the subset for a given intersection's specific turning direction. 80% of the subset is randomly selected as the input, and the left 20% is used to evaluate the performance of the model. Without loss of generality, we make a 10-fold cross validation for each subset. At the same time, we adopt the receiver operating characteristic (ROC) curve analysis and calculate the area under curve (AUC) value to assess different models' estimation effects. The results reveal that the proposed Boosting approach is practically promising in the field. In addition, the Boosting methodology is an open-ending one, which means any new promising models can be easily incorporated in the future.

2. The Boosting algorithm

In order to adapt the Boosting in our estimation research, we firstly give the original steps of Boosting algorithm for binary classification task. The Boosting algorithm is defined as follows (Freund et al., 1997):

Given: training set $(x_1, y_1),...,(x_m, y_m)$, where $x_i \in X, y_i \in Y = \{-1, +1\}$, T is the iteration number.

Initialize $D_1(i) = \{w_{11},...,w_{1m}\} = 1/m$, where $D_1(i)$ denotes the weight of this distribution on training example i on round t (here t=1).

For t=1,…,T

(1) Train weak learner using distribution D_t;

(2) Get weak hypothesis $h_t : X \to \{-1, +1\}$ with error:

$$\varepsilon_t = \Pr_{i \sim D_t}[h_t(x_i) \neq y_i] \tag{1}$$

(3) Choose $\alpha_t = \frac{1}{2}\ln(\frac{1-\varepsilon_t}{\varepsilon_t})$, here α_t denotes the final classifier's coefficient on round *t*.

(4) Update the weight's distribution:

$$D_{t+1}(i) = \frac{D_t(i)\exp(-\alpha_t y_i h_t(x_i))}{Z_t} \tag{2}$$

where Z_t is a normalization factor:

$$Z_t = \sum_{i=1}^{m} w_{ti} \exp(-\alpha_t y_i h_i(x_i)) \tag{3}$$

The final hypothesis is derived after *T* iterations which can be written as:

$$H(x) = sign(\sum_{t=1}^{T} \alpha_t h_t(x)) \tag{4}$$

This original Boosting classification algorithm can be easily modified for the regression purpose. In the second step of each iteration t, the threshold is added to the error calculation with the following criterion:

$$h_t : X \to \begin{Bmatrix} -1, |\, y_i - \tilde{y}_i \,| >= \varepsilon_{threshold} \\ +1, |\, y_i - \tilde{y}_i \,| < \varepsilon_{threshold} \end{Bmatrix} \tag{5}$$

where the $\tilde{y}_i$ denote the output result of hypothesis h_t on round *t*.

3. Experiment

3.1 Study area & Data pre-processing

The FCD data set contains 2636149 trajectories from about 20000 GPS-equipped taxicabs in Beijing collected by a business company from March to June in 2011, accounting for about 598 GB in data volume which leads a favorable coverage across the whole city of Beijing. We adopt the metropolitan road network of Beijing which is composed of 18857 nodes and 26621 road segments, holding an amount of 14614 intersections. To get a general idea of Beijing's congestion status, we choose 400 main intersections from all the 14614 ones, which mainly locate on the main roads, namely the expressways, the arterial streets and collector streets. The selected intersections in total represent the skeleton of Beijing's road network. The details are shown in the following Figure 1:

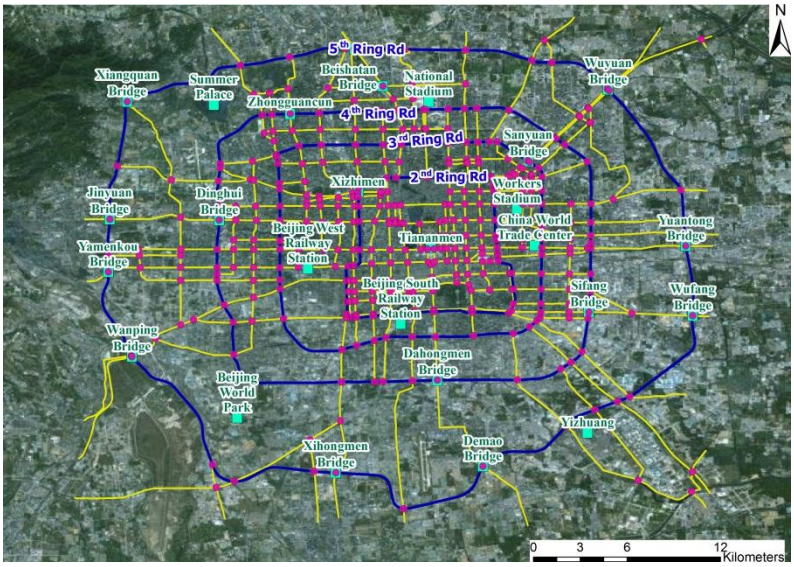

Figure1. Study area and main roads with 400 intersections

Each FCD record $FCD_i = \{gpsfile_1, gpsfile_2, ..., gpsfile_n\}$ contains a sequence of *GPS* log files, and represents the real travel log during a time period. The *GPS* log file $gps = \{pt_1, pt_2, ..., pt_n\}$ is a spatio-temporal point set with a series of single points $pt_i = \{lat_i, \mathrm{lo}n_i, timestamp_i\}$, indicating the exact latitude and longitude at the timestamp i.

Firstly, in order to extract the intersection delay records for a given intersection from the database, we utilize two data pre-processing procedures in this paper, namely the raw error filtering and the map-matching process. Secondly, the intersection delays for a specified intersection are calculated with a piecewise linear interpolation following Zhang et al.'s research result (Zhang et al., 2011). Thirdly, the intersection delay dataset for a given intersection is distributed among 96 time intervals (from 0:00 to 23:59 per 15 min during a day). In order to overcome the data sparseness and missing data problem caused by the unbalanced FCD spatio-temporal distribution, we employ the principal curve method to get a reliable estimation result of the missing value, which also represents the changing trend of the whole intersection delays. For more details, please refer to (Liu et al., 2013). Finally the total intersection delay dataset among the 400 main intersections is constructed. Each record contains a sequence of delays in each direction for a given intersection during 96 time periods in a day. The mean value of a time period is then taken as the representation of this time period's intersection delay. We also subdivide this dataset according to different weekdays (Monday to Sunday), because we believe that the delays' spatio-temporal characters of a given intersection may change along the weekdays.

3.2 The Boosting algorithm implementation

Six individual predictors including linear least squares regression(LLSR), autoregressive moving average (ARMA), historical mean (HM), Artificial Neural Network (ANN), Radical Basis Function Neural Network (RBF-NN), Support Vector Machine (SVM) are selected as the essential preparation of our research. These models have been proved to be effective in past studies. The first three methods belong to parametric techniques, while the others belong to nonparametric ones. Since the two types of techniques have

different capabilities to capture data characteristics in linear or nonlinear domain, we can provide comprehensive comparisons of the combined models composed of these components.

These six models are designed with the same input and output: we takes the first three time periods' intersection delay values as the input, while the fourth time period's intersection delay value as the output. For the three parametric techniques, the LLSR (*d*) model arises from the empirical observation that there exists a linear relationship between the first three time periods and the fourth time period. This observation leads to the prediction scheme by means of linear regression with time-varying coefficients. The ARMA(*p*,*q*) model is one of the most frequently used parametric models in time series forecasting, which is due to its flexibility in approximating many stationary processes and to its computational efficiency. In the HM model, *K* represents the number of the historical time periods (Here *K* is 3). Its results are obtained from the average of the three historical intersection delays. For the three nonparametric models (ANN, RBF-NN, and SVM), the topological structures and the input parameters play an important role. In order to determine the best parameters' combination, we use a simple grid search approach for each model's parameter selection. With this method, the ANN contains one hidden layer with 5 neurons, and the topological structure is 3-5-1. The initial weight and bias value for the ANN model are also determined. Specifically, we produce the RBF-NN with zero error on training vectors. It contains as many radial basis transfer function neurons as there are input vectors. The spread is set appropriate (five in the experiments) to ensure that the network function is smoother with better generalization for new input vectors occurring between input vectors used in the design and meanwhile to avoid each neuron being effectively responding in the same large area of the input space. For the SVM approach, the RBF kernel is selected and the appropriately adjusted parameters help to obtain better performance. With the grid search method, the best penalty parameter C and the RBF kernel parameter γ is selected. The six tuned models are then placed into the Boosting process which is descripted in Section 2. Here we set the $\varepsilon_{threshold}$ as 5 seconds.

3.3 Comparison methods and evaluation criteria

For comparison purpose, we also try some linear combining methodologies except for the six tuned models themselves. Four frequently used methods including equal weights method (Jose et al., 2008), optimal weights method (Granger et al, 1984), minimum error method (Yu et al., 2005), and minimum variance method (Lilian et al., 2000), are applied to determine the weights used in the combined forecasts.

In order to evaluate the model's quality, we first choose the subset for a given intersection's specific turning direction. 80% of the subset is randomly selected as the input dataset, and the left 20% is used to evaluate the performance of the model. Without loss of generality, we make a 10-fold cross validation for each subset. At the same time, we adopt the receiver operating characteristic (ROC) curve analysis and calculate the area under curve (AUC) value to assess different models' estimation effects.

3.4 The result

In our preliminary work, the six individual models' AUC values are among the range 0.65~0.77, which means the estimation effect is '*fair*' according to Swets' suggestion

(Swets, 1988). The HM approach performs the worst (0.65) while SVM the best (0.77). For the four linear combined models with different weight strategies, the models have a better performance than the individual ones, with the AUC value ranging from 0.85~0.91, which means the linear combined models are '*good*' (ranging from 0.85~0.90). One of this model using minimum variance method performs the best with the AUC value 0.91. But the fluctuation of this model during the peak time (7:00~9:00 and 17:00~20:00) still cannot be ignored. For the proposed Boosting approach, the AUC value reaches the crest of 0.95, which means that the proposed method is '*excellent*' to estimation the selected intersection delays during the 96 time period in different weekdays.

4. Conclusion

We propose a novel intersection delay estimation method based on Boosting. Compared with other six individual methods (LLSR, HM, ARMA, ANN, RBF-NN, SVM), this Boosting method performs the best. We also test our method with four other linear combined methods which are based on four different weight determination strategies: equal weights method, optimal weights method, minimum error method, and minimum variance method, the final result shows that our proposed Boosting method also outperforms these four approaches, with the AUC value 0.95. In addition, the Boosting methodology is an open-ending one, which means any new promising model can be easily incorporated. This method paves a new way into the intersection delay estimation research.

5. References

Aashtiani, H.Z., and Iravani, H. "Use of Intersection Delay Func-tions to Improve Reliability of Traffic Assignment Model." The 14th Annual International EMME/2 Conference, Chicago, IL, 1999.

Ban X J, Herring R, Hao P, Bayen A M (2009). Delay pattern estimation for signalized intersections using sampled travel times. Transport Res Rec, 2130(1): 109–119.

Ding, Z. A Static Traffic Assignment Model Combined with an Artificial Neural Network Delay Model. Ph.D dissertation, Department of Civil and Environmental Engineering, Florida International University, Miami, FL, 2007

Fangfang Zheng, Henk Van Zuylen (2010). Uncertainty and Predictability of Urban Link Travel Time. Transportation Research Record: Journal of the Transportation Research Board, Volume 2192(1), 136-146.

Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to Boosting. Journal of Computer and System Sciences55(1) (1997) 119–139

Granger C W J, Ramanathan R. Improved methods of forecasting. Journal of Forecasting 1984; 3 :197–204

Gühnemann A, Schäfer R P, Thiessenhusen K U, Wagner P. (2004). Monitoring traffic and emissions by floating car data. Institute of Transport Studies Working Paper, Issue ITS-WP-04 – 07, Sydney, Australia

Herring R, Hofleitner A, Abbeel P, Bayen A (2010). Estimating arterial traffic conditions using sparse probe data. In:Proceedings of the 13th International IEEE Conference on Intelligent Transportation Systems, September 19 –22, Madeira Island, Portugal,929– 936

Horowitz, A.J. "Intersection Delay in Region Wide Traffic Assign-ment: Implications of the 1994 Update of the Highway Capacity Manual." Transportation Research Record , No. 1572, Transportation Research Board of the National Academies, Washington, DC, (1997): 1–8.

Jose V R R, Winkler R L. Simple robust averages of forecasts: some empirical results. International Journal of Forecasting 2008; 24:163–169.

Kwong, K., Kavaler, R., Rajagopal, R., Varaiya, P. Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors[J]. Transportation Research Part C: Emerging Technologies,Vol.17(6):586-606. 2009

Lilian M de Menezesa, Derek W. Bunnb, James W Taylorb. Review of guidelines for the use of combined forecasts. European Journal of Operational Research 2000; 120:190–204.

Long J, Gao Z, Zhao X, Lian A, Orenstein P (2011). Urban traffic jam simulation based on the cell transmission model. Netw Spat Econ, 11(1): 43 – 64

Nielsen O A, Frederiksen R D, Simonsen N (1998). Using expert system rules to establish data for intersections and turns in road networks. Int Trans Oper Res, 5(6): 569–581

Oh, C. Anonymous Vehicle Tracking for Real-Time Traffic Performance Measures.PhD thesis. University of California, Irvine, 2003

Opitz, D.; Maclin, R. (1999). "Popular ensemble methods: An empirical study". Journal of Artificial Intelligence Research 11: 169–198. doi:10.1613/jair.614

Ritchie, S. G., S. Park, S. T. Jeng, and A. Tok. Anonymous Vehicle Tracking for Real-Time Freeway and Arterial Street Performance Measurement. Technical report. Research Report, UCB-ITS-PRR-2005-9, California PATH, Berkeley, 2005.

Sun L (2007). An approach for intersection delay estimate based on floating vehicles. Dissertation for Master Degree. Beijing :Beijing University of Technology(in Chinese)

Swets KA. Measuring the accuracy of diagnostic systems. Science, 1988, 240: 1285-1293.

Viti, F., and H. J. van Zuylen. Modeling Queues at Signalized Intersections. In Transportation Research Record: Journal of the Transportation Research Board, No. 1883, Transportation Research Board of the National Academies, Washington, D.C., 2004, pp. 68–77.

Winter S (2002). Modeling costs of turns in route planning. GeoInformatica, 6(4): 345-361

Xiliang Liu, Feng Lu, Hengcai Zhang, Peiyuan Qiu(2013). Intersection delay estimation from floating car data via principal curves: a case study on Beijing's road network. Frontiers of Earth Science, in press, DOI: 10.1007/s11707-012-0350-y

Yang Y. Combining forecasting procedures: some theoretical results. Econometric Theory 2004; 20:176–222.

Yu L, Wang S, Lai KK. A novel nonlinear ensemble forecasting model incorporating GLAR and ANN for foreign exchange rates. Computers & Operations Research 2005; 32(10):2523–2541

Zhang H C, Lu F, Zhou L, Duan Y Y(2011). Computing turn delay in city road network with GPS collected trajectories. In: Proceedings of the 2011 International Workshop on Trajectory Data Mining and Analysis. September 17-21, Beijing, China, ACM, 45–52

Zhao, Minyue, Li, Xiang (Corresponding author), 2013, Deriving Average Delay of Traffic Flow around Intersections from Vehicle Trajectory Data, Frontiers of Earth Science, in press.

Z. Zhou, Ensemble Learning. ;In Proceedings of Encyclopedia of Biometrics. 2009, 2.