2012 15th International IEEE Conference on Intelligent Transportation Systems
Anchorage, Alaska, USA, September 16-19, 2012

A Short-term Freeway Traffic Flow Prediction Method Based on Road Section Traffic Flow Structure Pattern

Ping Zhang, Kunqing Xie, Guojie Song

Abstract—**Accurate short-term traffic flow prediction is the foundation of the efficient and proactive management of freeway networks, especially on the abnormal traffic states. The relationship between traffic flow on the current section and the upstream stations can be used for predicting short-term traffic flow. In this paper, we reveal this relationship by the traffic flow structure pattern. The structure pattern can be drawn from real freeway toll data and a few video detective cameras on the freeway segments. Based on the stability pattern, a new traffic flow prediction algorithm has been proposed. Experimental based on real data showed that the prediction method based on structure pattern is an effective approach for traffic flow prediction, especially on the abnormal traffic state.**

I. Introduction

Accurate short-term traffic flow prediction is the basis of the management of freeway networks. It can help with reducing the mean travelling time and alleviating traffic congestion. The key to the traffic flow prediction lies in how to utilize the real-time traffic data efficiently to predict the traffic status, including volume, velocity, etc., during the next period of time (usually 5-15 minutes).

Researchers all over the world have employed all kinds of methods for short-term freeway traffic flow forecasting. Addressing the problem of traffic flow prediction, Smith et al. compared nonparametric regression models and the autoregressive integrated moving average (ARIMA), and pointed out that the nonparametric regression models could be refined much further [1]. The experiments conducted by Sun et al which were based on real data demonstrate that LWL, compared with K nearest, can enhance the accuracy and the efficiency [2].

From the Wavelet Transform (WT) theory, statistical characteristic of short time traffic flow shows that the time series of the traffic data can be regards as normal and abnormal components [3]. The normal component represents the main trend of traffic flow, while the abnormal parts usually presents the atypical traffic flow, which makes the actual traffic flow fluctuate nearby the general trend. The vast majority of the academic studies on short-term traffic prediction mainly focus on normal traffic states. These methods can get good accuracy at such states. However, the dealt with typical or non-incident traffic conditions limits these studies applicability during abnormal traffic condition, such as traffic accidents, jam or other interference [4].

Recently, some prediction models of traffic flow during abnormal conditions have been developed. Such as an Online-Support Vector Regression model was used for traffic flow prediction on holidays or days with traffic incidents [5]. Neural network based methods have been used to predict travel times under abnormal weather [6]. But such black-box learning methods cannot explain the phenomenon of abnormal traffic from the essence, and another drawback of the existing methods is they do not careful analysis of the relation between traffic flow of the current section and its upstream stations.

In a freeway network, toll data can be obtained easily. The freeway toll data includes the information about where and when vehicles pass in and out of the freeway network, which, to a certain degree, reflects the vehicles' spatio-temporal statuses. The traffic Origin-Destination data can be directly obtained from such data. And with the help of a few video detective cameras, we can extract the road section traffic flow structure pattern, which reveal the fundamental spatio-temporal relationship between traffic flow on the current road section and the upstream stations. This information can effectively support the traffic volume prediction method, especially under abnormal traffic condition.

The main contributions of this paper are: 1) Revealing the freeway section traffic flow stability pattern by using the real freeway toll data and a few video detective camera on the freeway sections; 2) Based on the stability pattern, a new traffic flow prediction algorithm has been proposed; 3) Experimental based on real data show that proposed algorithm is reasonable and feasible, and that the accuracy is improved under abnormal traffic condition.

The rest paragraphs are organized as follows. The second part gives the structural analysis of freeway road section traffic flow. Section three talks about the prediction methodology via traffic flow structure pattern. Experimental studies are implemented in the fourth part. The last section gives a conclusion.

II. Structural Analysis of Freeway Traffic Flow

Traffic flow on a certain road section of freeway is affected by flow of the upstream stations; the relationship between them can reveal the essence of some abnormal traffic condition, like the section congestion caused by the upstream traffic surge. And this relationship is actually the

Ping Zhang is with Key Laboratory of Machine Perception (Minister of Education), Peking University, 100871, Beijing, China (e-mail: zhangping@cis.pku.edu.cn).

Kunqing Xie is with Key Laboratory of Machine Perception (Minister of Education), Peking University, 100871, Beijing, China (e-mail: kunqing@cis.pku.edu.cn).

Guojie Song is with Key Laboratory of Machine Perception (Minister of Education), Peking University, 100871, Beijing, China (e-mail: gjsong@pku.edu.cn).

traffic flow structure pattern. With the help of freeway toll data and a few video detective cameras, we can find this pattern, and use it to improve the accuracy of traffic flow prediction method.

A. Data Description

In order to measure the traffic flow structure stability pattern, we analyzed the freeway toll data of one province of China throughout December of 2010, which has a total number of 137 toll stations and 340 road sections (between adjacent stations),some of the sections are installed cameras for tracking cars. And 11000000 records of toll data are collected in this month. The details of toll data are shown in Table I.

TABLE I. MAIN PROPERTIES OF THE FREEWAY TOLL DATA

Name	Content
CarID	Car identifier
Sin	Origin station ID
Sout	Destination station ID
Tin	Time of the car enter freeway
Tout	Time of the car exit freeway

B. Traffic Flow Structure Pattern Analysis

Through toll records and cameras' data, we can generate a trajectory of the car. Then we can get the time that the car enter/exit a certain section and total section flow at a certain time interval. The time granularity is set to be 15 minutes.

The trajectory of a car can be represented as following:

$$\text{Traj} = \{(T_{Rin}, T_{Rout})_i\}(\text{i} \geq 1) \quad (1)$$

Where *Traj* means that a trajectory of a car is made by a consecutive road sections, T_{Rin} means the time of the car enter the road section *i*, and T_{Rout} means the exit time. After that, we can group the cars by the number of sections that those cars have passed. Then the structure of section traffic flow at current time interval S_R can be expressed as a vector S_R, i.e:

$$S_R = < C_{G1}, C_{G2}, \dots, C_{Gn} > \quad (2)$$

Where, C_{Gi} means the count of cars on current road section at current time interval and those cars have just passed *i* road sections. The sum of C_{Gi} is the traffic flow of this section, say Q_R. On the other hand, the component C_{Gi} of the traffic flow structure also represents the relationship between current road section and its upstream station which is *i* sections far away.

To find out the structure pattern of traffic flow on a certain road section, we first show that only a small set of component C_{Gi} is necessary for reasonably accurate construction the relationship between section and its upstream stations.

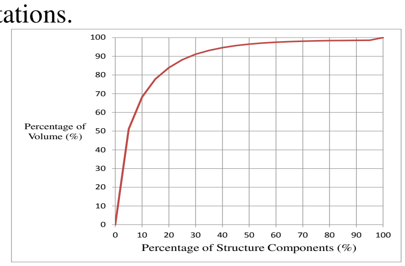

Figure 1. Number of components that constitute each section flow (Total 20 componets)

In Figure 1, we plot the cumulative distribution function (CDF) in a month of each section flow structure. Figure shows that 83% volume of any given section is composed of no more than 4 (17%) significant components, and generally many fewer. This surprising result means that we can think of each section flow as having only a small set of "features." Then, the volume contributed by 3 significant components average a month is summarized in Figure 2.

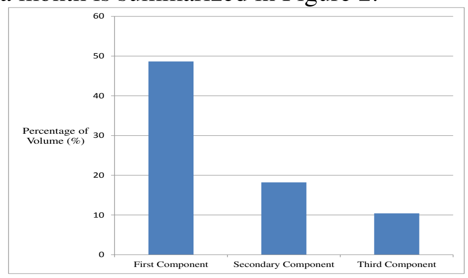

Figure 2. Volume percentage of each significant component contribute average a month

The Figure 2 shows the surprising result that the vast majority of traffic volume is contributed by the first component. Thus, we can draw a conclusion that the contribution of section volume by each component is uneven, and is mainly concentrated in top 3 significant components. That is to say, to get a structure pattern of a section traffic flow is to find these 3 significant components. Thus, we can find the principle upstream stations of current road section.

Figure 3 shows the average sum percentage of top 3 significant components among all the sections in a week, as long as with the average section flow.

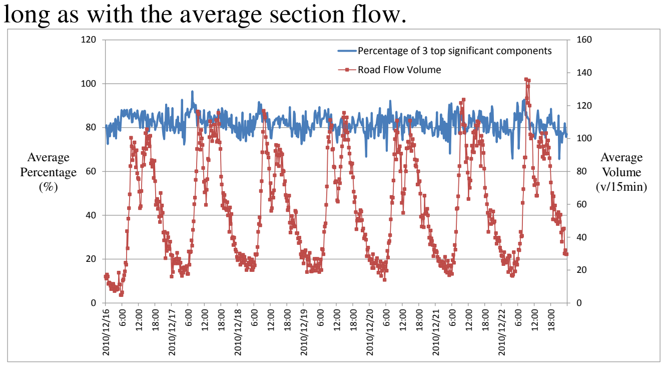

Figure 3. Sumpercentage of top 3 significant components contribute average a week and averagesection flow

From this figure we can see, there is a kind of periodicity of the section flow, while the percentage of 3 top significant components has a smaller fluctuations than the flow. That is to say, regardless the flow change, the contribution ratio of top 3 principle components is always keep at a certain degree, say about 80%.To further verify the stability of the pattern, we use Coefficient of Variation(CV) as stability index [7].

Coefficient of Variation is the percentage variation in mean, standard deviation being considered as the total variation in the mean. The series of data for which the coefficient of variation is large indicates that the group is more variable and it is less stable or less uniform. If a coefficient of variation is small it indicates that the group is less variable and it is more stable or more uniform.

$$\mathrm{CV} = \frac{1}{\overline{\mathrm{X}}}\sqrt{\frac{1}{\mathrm{N}}\sum_{\mathrm{i}=1}^{\mathrm{N}}(\mathrm{X_i} - \overline{\mathrm{X}})^2} \tag{3}$$

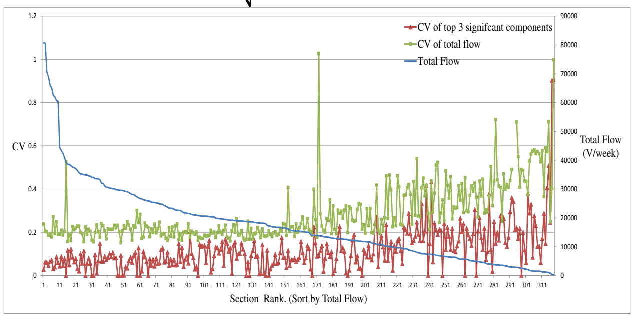

Figure 4. CV of percentage of top 3 significant components and section flow of every section (sort by total section flow)

Figure 4 shows the CV of flow structure pattern (i.e.: sum percentage of top 3 significant) is more stable than section flow. And the stability of high flow sections is better than low flow sections, both flow and flow structure pattern. Then conclusions can be drawn from the foregoing analysis:

- Through the trajectory of cars we define the structure of section traffic flow, and the contribution of each component is uneven, about 83% volume of any given section is composed of less than 4 significant components.
- There is a kind of periodicity of the section flow.
- Flow structure pattern is more stable than section flow.

III. Prediction Method Based on Traffic Flow Structure Pattern

Like weather/ hydrologic forecasting, traffic prediction needs not only the data a moment ago, but also relevant knowledge in the field of traffic. The structure pattern of traffic flow is just the domain knowledge which will enhance the prediction methods' accuracy, especially in the abnormal traffic state.

A. *Locally Weighted Learning*

This paper focuses on improve the lazy learning. Lazy learning is a learning method in which generalization beyond the training data is delayed until a query is made to the system. A typical lazy learning method is locally weighted learning (LWL). Locally Weighted Learning, which is a Lazy Learning Method based on local models, and which employs Local Linear/ Nonlinear Models to fit the neighbor points and then uses their values to predict the values of the query points. Besides, LWL can add currently acquired new data into the historical database to help with learning. This algorithm can be easily realized without any beforehand-given model, and can achieve a high accuracy of prediction [8].

For the locally weighted learning, lazy behavior is based on abundant training set. Unfortunately, in some cases, this knowledge is poor to support the learning process due to the inadequate history data and new queries. Such as abnormal traffic state, traditional locally weighted learning will lose its ability and some strategies should be introduced to assistant regression [9].

Therefore, by analyzing the structure pattern of road section traffic flow, we reveal the fundamental spatio-temporal relationship between traffic flow on the current section and its upstream stations. And then adds this domain knowledge into the regular LWL prediction method, and utilizes traffic system's inherent characteristics to help the learning algorithm to choose features and assistant regression, so that the accuracy of prediction and the efficiency of algorithm could be enhanced.

B. *Prediction Method Based on Traffic Flow Structure Pattern*

The fundamental thought of improving LWL with the stability structure pattern of traffic flow is, on the one hand, to direct LWL to choose features for the feature space, and on the other hand, to use the upstream stations entrance flow to correct the prediction of the section traffic flow:

- **Inherent periodicity:** Because the section flow have good inherent periodicity, especially those high flow sections, their volumes at previous hours can be chosen as features;
- **The feature upstream stations:** The structure pattern shows that section flow is composed of less than 4 significant components. That is to say, only a few upstream stations have an influence on current section. these upstream stations entrance flow can be chosen as the feature of LWL
- **The correction of the feature Os' volumes:** Unless abnormal points occurred in history, it is hard, when new abnormal points occur, for the prediction algorithm to response to an abnormal status because only the section own volumes are taken into account as the feature values for the regression analysis, so the current flow of the upstream stations relevant to the current section are also needed for the correction of the prediction value; Table. II gives the explanation of the variables used in Algorithm

TABLE II. Variables Description

Name	Content
Q_R	Traffic flow of the target section
Q_{up}	Feature upstream station entrance flow, the station is chosen by sturcture pattern
x_i	Data point, is a vector, select previous time Q_R, Q_{up} and current time *t* as its features. $<Q_{R(t-5)},\ldots Q_{R(t-1)},\ Q_{up1}\ldots,Q_{up3},\ t>$
y_i	Traffic flow of the target section correspoding to the x_i
D	Labelled training data, where $\mathrm{D} = \{(\mathrm{x}_i, y_i) \mid i = 1,2,\ldots,n\}$ (4)
x_q	Query point, which is the position where we want a prediction$\hat{y}_q$
$d(\mathrm{x}_i,\mathrm{x}_q)$	Measures the relevance of training points for the current prediction, $d(\mathrm{x}_i,\mathrm{x}_q) = \sqrt{\sum_j(\mathrm{x}_i - \mathrm{x}_q)^2}$ (5)
$K(\mathrm{d})$	Computes for each distance value a correspondingweight, a Gaussian Kernel here $K(\mathrm{d}) = \exp(-d^2)$ (6)

$V(t)$	Give the predictionvalue based on $K(\mathrm{d})$, where $$V(t)=\frac{\sum \mathrm{y}_i K\left(d(\mathrm{x}_i,\mathrm{x}_q)\right)}{\sum K\left(d(\mathrm{x}_i,\mathrm{x}_q)\right)} \quad (7)$$
C_{Gs}	significant components in flow sturcture pattern
$C(t)$	Correction function based on structure pattern. $$C(t)=\frac{\sum_j Q_{upj} * w_{rj}}{P} \quad (8)$$ Where Q_{upj} is the flow at time $(\mathrm{t}-\Delta\mathrm{t})$, $\Delta\mathrm{t}$ is the transfer time from the station to the section, which is caculated by history average
P	The history average sum percentage of 3 top significant components $$P=\frac{\sum_j C_{Gsj}}{Q_R} \quad (9)$$
w_r	The history average percentage of the upstream stations entrances flow contribute to the current section. Which is drawn from history dataset

Algorithm: Prediction Method Based on Traffic Flow Structure Pattern

Inputs:
(1)Labeled training data D
(2)A sequence of given queries $\{\mathrm{x}_\mathrm{q}\}$
Outputs:
The output prediction vector $\hat{\mathrm{y}}_\mathrm{q}$of the given queries
Algorithm:
1. Initialization:
(1) Get the target section flow structure pattern
(2) Choose the feature upstream stations based on structure pattern
(3) Calculate the history average transfer time $\Delta\mathrm{t}$ and contribution rate w_r of each feature upstream stations.
2.Regression:
For each query x_q
- **For** each history point$(\mathrm{x}_i,\mathrm{y}_i)$
  Calculate $d(\mathrm{x}_i,\mathrm{x}_q)$
  End for
- Sort (d) and find 10 neighboring points in the space, then rank them in descending order according to distance.
- Get each 10 points weighted based on Gaussian Kernel $K(\mathrm{d})$
- Calculated $V(t)$
- Calculated $C(t)$
- Predict $\hat{y}_q$, where
  $$\hat{y}_q=\alpha * V(t)+\beta * C(t) \quad (10)$$
  $\alpha+\beta=1$, and initially $\alpha=\beta$. If currently the flow of target section has a very stable periodic pattern, enlarge the proportion of α, and vice versa
- Add the newly obtained points into the historical database of pattern.

End for

IV. Case Study

A. Data Description

The experiment dataset was composed of volume data of 21 consecutive days, December 2 to 22; year of 2010.We chose two high volume sections to take experiments. The target sections we concerned were highlighted by different colors in Figure 5. And we chose red one as the typical section.

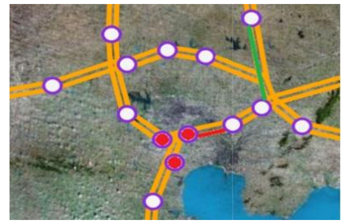

Figure 5. Data source for experiments

B. Traffic Flow Structure Pattern Analysis

At first, we will analysis the traffic flow structure pattern of typical section. Figure 6 gives the section volume through 21 days; apparently, there is 2 days' flow much higher than the regular days. That's why we chose this section as typical, to test our algorithm accuracy during the traffic abnormal state. And the abnormal day is day 19 and day 22.

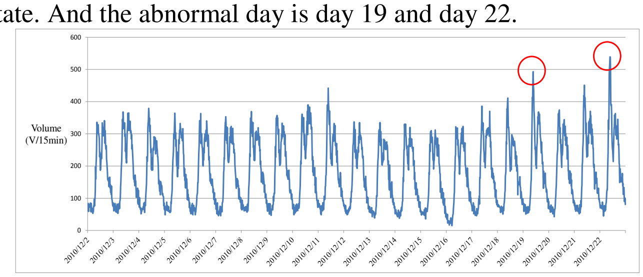

Figure 6. Target section flow during 21 days (red section)

Then we want to find the flow structure pattern, Figure 7 shows the percentage of each component contribute to the flow, it clearly illustrates that the principle components is component No.0 and component No.2，thus, we can find the feature upstream stations, which are marked by red also in Figure 5.

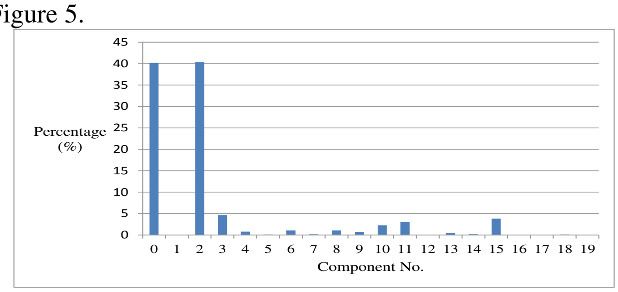

Figure 7. Volume percentage of each componentcontribute average 21 days (red section)

To further test the structure pattern stability, we plot the sum percentage of component No.0 and component No.2 in Figure 8, it shows that the structure pattern has a good stability even when the abnormal flow happened both in day 19 and day 22, and the Coefficient of Variation of this data series is 0.06066029.

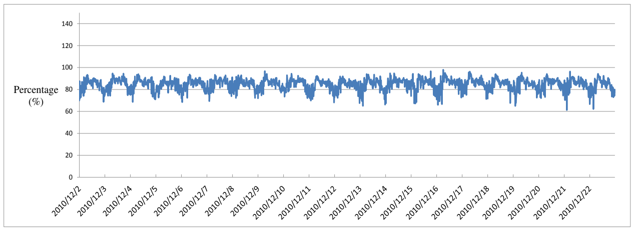

Figure 8. Sum percentage of top 2 significant components in 21 days (red section)

Now we find the relationship between the section and upstream stations, Figure 9 gives the total entrance volume of those feature stations. We can find that just because the increasing of those feature stations' flow, the target section flow is abnormal at day 19 and day 22.

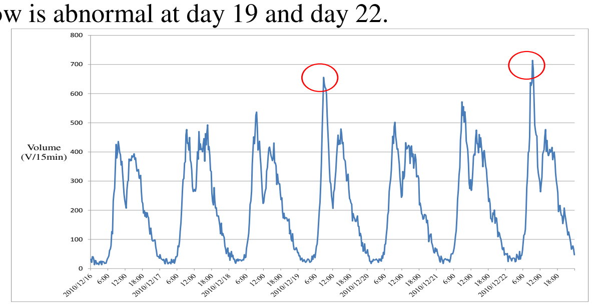

Figure 9. Sum flow of the 3 feature stations in last 7 days (red section)

C. Experiment results and analysis

The first 14 days were used for training the prediction model, the latter seven days was for prediction to test the model's performance. The regular LWL method which is just use the section's periodicity $V(t)$ in the regression was adopted as the comparison method. Our method referred as *SP*, regular LWL as *LWL*

Root Mean Square Error (RMSE) was used to estimate the accuracy of the prediction, defined as.

$$\mathrm{RMSE}(y, y') = \sqrt{\frac{\sum_{n=1}^{N}(y(n) - y'(n))^2}{N}} \quad (11)$$

Where y is the real volume and y' is the corresponding prediction while N is the total number of prediction values. The model has smaller RMSE is better accuracy. The experiments results are below:

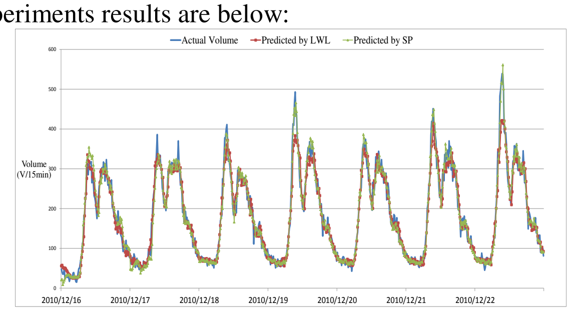

Figure 10. Prediction result of the whole week (red section)

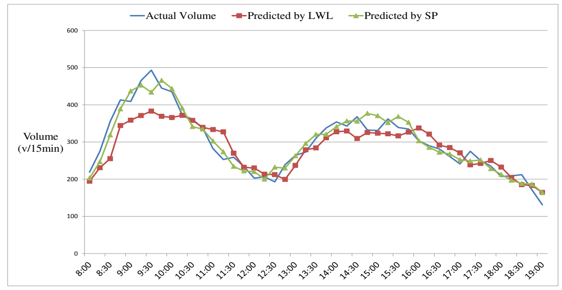

Figure 11. Prediction result of day 19 (8:00—19:00,red section)

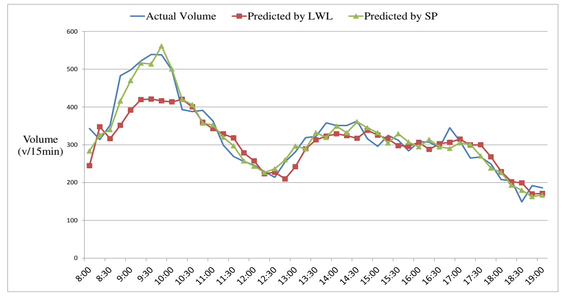

Figure 12. Prediction result of day 22 (8:00—19:00, red section)

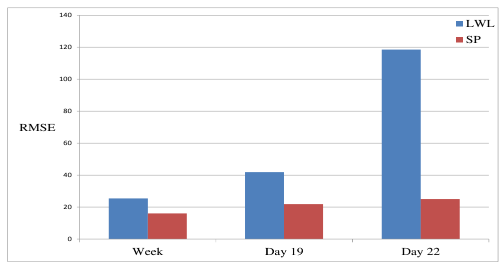

Figure 13. RMSE of each experiments (red section)

TABLE III. RMSE OF EACH EXPERIMENTS (RED SECTION)

RMSE	LWL	SP
Week	25.44596	16.05697
Day 19	41.90131	21.87065
Day 22	118.517	25.03003

From the figure and table, we can easily find 1) our method is better than the regular LWL method in the overall week. 2) In the abnormal days, such periodicity alone are not enough for the regression to capture an abnormal status, but on the contrary, when comes to the method based on the structure pattern, the section's periodicity are combined with the correction according to the relevant upstream stations entrance flows, the prediction algorithm can immediately response to the abnormal status. That is to say, the domain knowledge found by traffic flow structure pattern explains the phenomenon of abnormal traffic from the essence: the abnormal flow on the section is just because there is a sudden increasing in the entrance flow of the upstream station.

Another section's (green) experiments are plotted below:

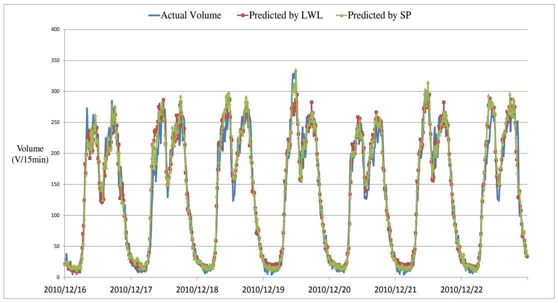

Figure 14. Prediction result of the whole week (green section)

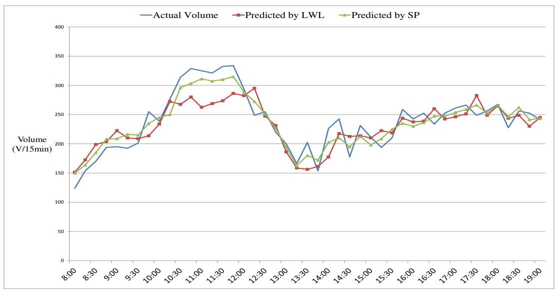

Figure 15. Prediction result of day 19 (green section)

TABLE IV. RMSE OF EACH EXPERIMENTS (GREEN SECTION)

RMSE	LWL	SP
Week	21.68704971	19.24174417
Day 19	28.2160563	16.37902792

Also, the result shows that our method outperforms the regular LWL method, especially on the day 19 which have an atypical flow peak.

In order to verify our method, we choose regular artificial neural networks (ANN) as comparison. The input layer of ANN is last 4 previous section flow (now, 15min, 30min, 45min) and the output layer is next 15min's section flow. The experiment was taken on the green section.

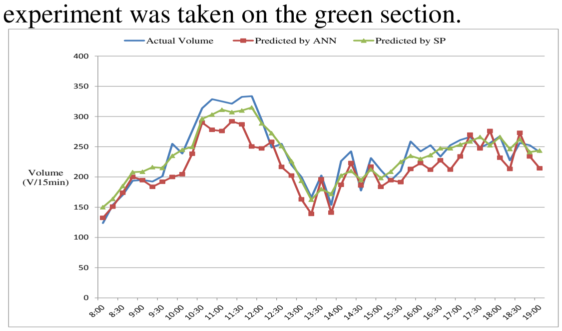

Figure 16. Prediction result of day 19 (green section)

TABLE V. RMSE OF EACH EXPERIMENTS (GREEN SECTION)

RMSE	LWL	SP
Week	23.20474221	19.24174417
Day 19	30.07704768	16.37902792

In the figure 16 and Table V, the result shows that our method outperforms the regular ANN method.

V. CONCLUSION

Traffic flow of the current section has great correlations with flow of its upstream stations. We analyzed this relationship by using the structure pattern of traffic flow. Then traffic flow of the current section at next time point was predicted by using current traffic flow and upstream stations entrance flow information via LWL. Experiments were carried out on the data of one Province of China. At first, we construct the structure pattern of traffic flow, and then with this pattern we find the principal entrance stations of all the sections, from which we choose 2 sections as targets, with the pattern information and LWL, we give a prediction of those sections' flow in a week. The results showed that the prediction method based on structure pattern is an effective approach for traffic flow prediction, especially on the abnormal traffic state which is just caused by upstream stations' flow sharply increased.

REFERENCES

[1] Brian L. Smith, Billy M. Williams, R. Keith Oswald. "Comparison of parametric and nonparametric models for traffic flow forecasting". *Transportation Research Part C*, 2002, 3(1): 303-321.

[2] Yijun Sun, Dapeng Wu."Feature extraction through local learning",*Statistical Analysis and Data Mining archive*, 2009,2(1):34-52

[3] Yuesheng Gu, Yancui Li, Jiucheng Xu and Yanpei Liu, "Novel Model Based on Wavelet Transform and GA-fuzzy Neural Network Applied to Short Time Traffic Flow Prediction" ,*Information Technology Journal,*2011,10: 2105-2111.

[4] Fangce Guo, John W. Polak and Rajesh Krishnanm. "Comparison of modeling approaches for short term traffic prediction under normal and abnormal conditions", *13th International IEEE Conference on Intelligent Transportation Systems (ITSC'10)* 2010, pp1209 - 1214

[5] S. Lam and T. Toam, "Short-term travel time prediction using suppor vector regression," *Proceedings of the 87th Annual Meet Transportation Research Board*, Washington D.C, USA,2008

[6] Y. Tao, F. Yang, Z. J. Qiu, and B. Ran, "Travel time prediction in the presence of traffic incidents using different types of neural networks, "*Proceedings of the 85th Annual Meeting of the Transportation Research Board*", Washington D.C, USA,2005/

[7] C. Jin, D. X. Wei, and S. H. Low. "FAST TCP:Motivation, architecture, algorithms, performance.", *In IEEE INFOCOM 2004*, Hong Kong, March 2004

[8] Christopher G. Atkeson, Andrew W. Moore, Stefan Schaal. "Locally weighted learning for control", *Artificial Intelligence Review 11*, 1997, 2(1): 75–113.

[9] Lei Han, Jianying Wu, Ping Gu, Kunqing Xie, Guojie Song, Shiwei Tang, Dongqing Yang. "Adaptive Knowledge Transfer Based on Locally Weighted Learning.",*The 2010 Conference on Technologies and Applications of Artificial Intelligence (TAAI 2010)*,Taiwan, China, 392-397, 2010.