Fluctuation Similarity Modeling for Traffic Flow Time Series: A Clustering Approach

Shan Jiang, Shuofeng Wang, Xin Pei
Department of Automation
Tsinghua University
Beijing, China 100084

Zhiheng Li
Graduate School at Shenzhen
Tsinghua University
Shenzhen, China 518055
zhhli@tsinghua.edu.cn

Guo Weiwei
Beijing Key Laboratory of Urban Road
Intelligent Traffic Control
North China University of Technology
Beijing, China 100144

***Abstract*—Traffic time series analysis is important because of its use in traffic control and travel time prediction. In this paper, we discuss how to cluster traffic time series that have similar fluctuation patterns. We use simple average detrending method and only study the residual time series. Second, we use principle component analysis (PCA) on raw data and use the weight of the first d-components as the features of the time series. Third, we use k-means algorithm to cluster the traffic time series. Finally, we study the results of the clustering algorithm and discuss the origins of the clusters. In summary, the most important factors of clustering results are urban/rural area, direction and in/not in ramp entrance.**

Keywords—traffic time series; fluctuation pattern; detrending; PCA; k-means

I. INTRODUCTION

Traffic time series analysis has been studied from different aspects during last three decades [1]-[5]. It had been shown in some papers [6][7] that traffic time series at different space locations may have causality relations [8][9][10].

In this paper, we focus on finding nodes that have similar fluctuation patterns. The topic of time series clustering has received continues interests in many research fields, such as engineering, science, finance, economic and government [11]. Such clustering method is also useful in transportation studies, since we can use the obtained results to build better traffic prediction model and missing traffic data imputation model.

Researchers generally assume that traffic time series exist day-invariant trend which can be separated with the reminding residual time series by detrending method. The residual time series represent the short-term fluctuation component of the time series. The detrending method can help to build more accurate prediction model and missing data imputation model. Detrending method also benefits to analyze the essential correlation between traffic time series record at different locations [12]. In this paper, our purpose of clustering is to find the traffic time series that have similar fluctuation pattern, so we first identify the daily trend of the traffic flow time series and only study the residual time series. There exist many trend definitions, such as simple average daily trend [12], wavelet based trend [13][14]. We will use the simple average daily trend since it is simple and introduces the least influence on fluctuation patterns.

Second, we use dimensionality reduction method to retrieve the most salient features of traffic flow time series. There are two main methods for dimensionality reduction: PCA [15][16] and ICA [17]. The ICA assumes the observed signals are linear mixture of non-gaussian and mutual independent source signals. Because the residual time series may obey Gaussian distribution and PCA is more suited to the problem about traffic time series[18], we use PCA method.

Third, for each time series, we use the weights of the first d principle components as the features of the time series. We generally use k-means [19][20] and GMM [21] to cluster the time series. We compare these two methods and determine adopt k-means algorithm for its simplicity.

In the rest part of this paper, we will first present the clustering methods. Second, we will analyze the results of the clustering algorithm and discuss which nodes have similar fluctuation patterns. Third, we will discuss the origins of the clusters and the applications of our method.

II. METHOD

A. Simple average trend

Traffic time series have similar daily trends in continuous days. Generally, this trend is like an M-shape, whose first peak corresponds to the morning rush hour and second peak corresponds to the evening rush hour [12]. The trend does not reflect the fluctuation pattern. So, in this paper, we use the simple average daily detrending method to remove the trend [12].

Suppose the traffic time series of N consecutive day can be written as N one-dimensional vectors as follow

$$\vec{y}_1 = [y_{1,1}, y_{2,1} \cdots y_{n,1}] \quad (1)$$

$$\vdots$$

$$\vec{y}_N = [y_{1,N}, y_{2,N} \cdots y_{n,N}] \quad (2)$$

In (1) (2), n is the number of sample point per day. In this paper, our sample time interval is set as 5 minutes, so the n is 288. The daily trend is defined as

$$\vec{y}_{trend} = [\sum_{i=1}^{N} y_{1,i}, \sum_{i=1}^{N} y_{2,i}, \cdots \sum_{i=1}^{N} y_{n,i}] \quad (3)$$

978-1-4673-6596-3/15 $31.00 © 2015 IEEE
DOI 10.1109/ITSC.2015.143

The residual time series of i th day is defined as

$$\vec{y}_i = \vec{y}_i - \vec{y}_{trend} \tag{4}$$

The original time series, daily trend and residual time series are shown as Fig.1. In the rest of the paper, we will always use the residual time series.

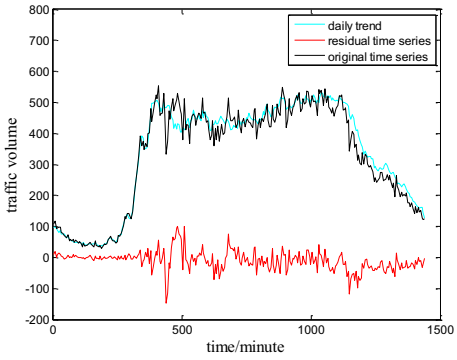

Fig. 1. An example of original time series, daily trend and residual time series

After we get all days' residual time series, we can merge them to raw data $Y \in R^{m,s}$, m is the number of time point in all days, s is the number of nodes. $Y_{i,j}$ represents the residual time series in i th time point at node j.

B. PCA based dimensionality reduction

Principal Component Analysis [15][16] is a classical dimensionality reduction method. The basic idea of PCA is projecting high-dimension variable into an orthogonal space in which the first d coordinates have larger variance.

The row data is as $Y \in R^{m,s}$, $Y_{i,j}$ represents the residual time series in i th time point at node j. According to PCA assumption, there exists a latent matrix $X \in R^{m,s}$ which is the projection of the Y in the space of principle components. Y and X satisfy

$$Y = XP + M \tag{5}$$

where the $M \in R^{m,s}$ is the mean shift, and the $P \in R^{s,s}$ is the projection matrix [9][10][16]. It needs to note that the $P_{i,j}$ is the j th node's i th component's weight. The algorithm of estimating the projection matrix P can be find in [16]. We select the first d components and use the weight of each component to cluster the nodes.

C. Cluster method

We can cluster all nodes by using k-means [19][20]. The features of the nodes are the weights of the first d components. When we have extracted the features of all traffic time series, we get a data set $W = \{w_1, w_2 ... w_s\}$ which has s nodes, $w \in R^d$ are the first d components' weights of traffic time series. Our aim is to partition the data set W into K clusters. The objective function of the k-means is generally defined as

$$\min_{S_k, \mu_k} J = \sum_{k=1}^{K} \sum_{w_i \in S_k} \| w_i - \mu_k \|_2^2 \tag{6}$$

where u_k is the k th cluster centroid, S_k is the data set that consisting of the nodes in k th cluster. It needs to note there exist many different objective function, we adopt this most common objective function.

We use a coordinate descend method to solve the k-means problem. First, we assign each node to the class centroid which is closest to it. By this way, we can obtain K new clusters. Second, we update the new centroids of the K new clusters. Finally, we can obtain the clustering results by repeated iteration of this two steps.

III. Data Analysis

A. data and settings

All the data used in this paper are extracted from the publicly accessible PeMS dataset [23]. The traffic flow time series are recorded at 1000 freeway traffic flow stations during August 1, 2011 to August 31, 2011. The sample interval of the raw data is 5 minutes. When using PCA, we select the first 6 components. In next chapter, we first compare the results of the k-means and GMM.

B. The cluster numbers

We use RSS (root sum of squared error) as the evaluation criteria of the k-means with each K parameter. The RSS can be defined as

$$RSS = \sqrt{\sum_{k=1}^{K} \sum_{w_i \in S_k} \| w_i - \mu_k \|_2^2} \tag{7}$$

where u_k is the k th cluster centroid, S_k is the data set that consisting of the nodes in k th cluster, the K is the number of cluster. The RSS with each K parameter is shown as table 1.

TABLE 1. The RSS of K-means with each K parameter

K parameter	RSS	K parameter	RSS
1	2.2706	5	1.7319
2	2.0936	6	1.6620
3	1.9607	7	1.5926
4	1.8493	8	1.5446

We have checked the cluster results with different k parameter. The cluster results of k=5 or k=6 are almost same as the cluster results of k=4 except few nodes (less than 5). Because of the space limitation, we will not show all these results as figure.

According to the numerical results of table 1 and cluster results, we choose a proper number of cluster as 4. The clustering results are shown as Fig. 2 in which different color means different class label.

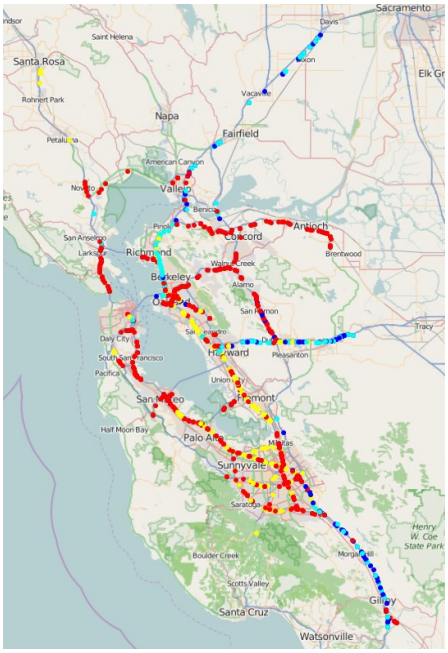

Fig. 2. Clustering results

C. *Clustering results*

After using the PCA, the first 6 components are shown as Fig. 3. The amplitudes of the first 2 components are larger than those of the rest components. This indicates there exist some sudden high traffic flow demands. The latter components mainly represent periodic random fluctuations. The weights of the first 6 components for the 4 class centroids are shown in Fig.4, corresponding to the red, cyan, yellow, blue nodes in Fig.2. The 6 color bars of each class represent the weights of 6 components.

In Fig.4, we find the weights of 4 class centroids are significantly difference. This means the fluctuation patterns of 4 classes are different. To illustrate the differences of 4 classes, we choose 4 time series whose weights are closest to 4 class centroids and plot their residual time series in Fig. 5.

To demonstrate that the nodes come from same class have similar fluctuation patterns, we randomly choose five nodes from the 3rd class (yellow color nodes) and plot their residual time series in Fig. 6. We can find a similar fluctuation pattern in these time series.

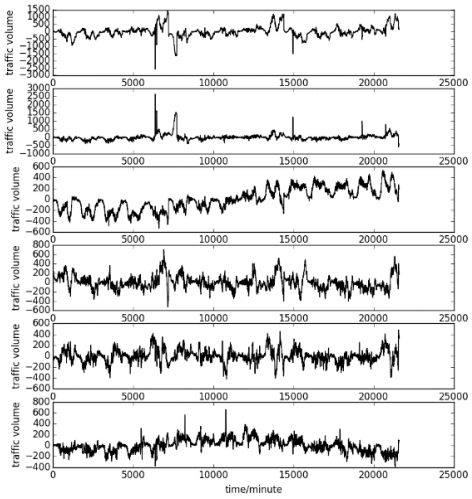

Fig. 3. The first 6 principle components

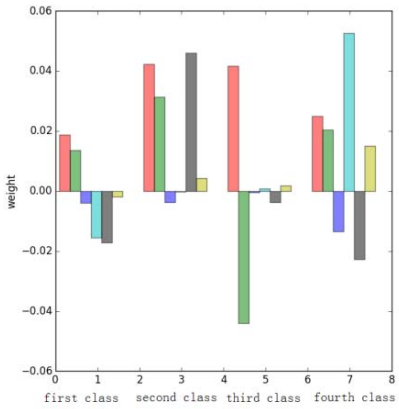

Fig. 4. The bar chart of each class centroid, the four classes corresponding to the red, cyan, yellow, blue nodes in figure 4. The 6 color bars represent the weights of 6 components.

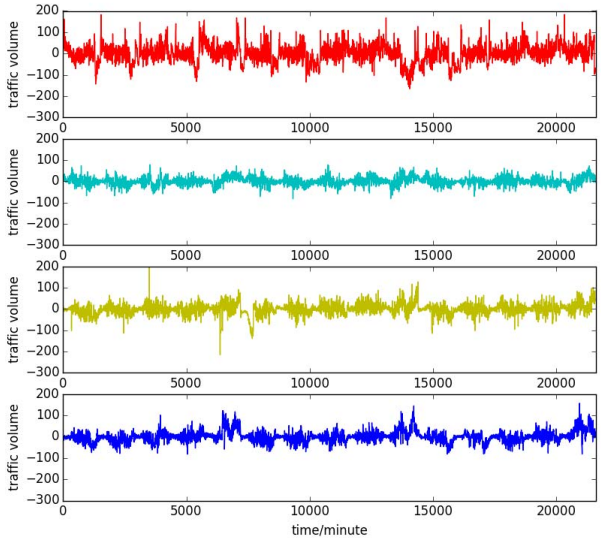

Fig. 5.The residual time series which are closest to each class centroid

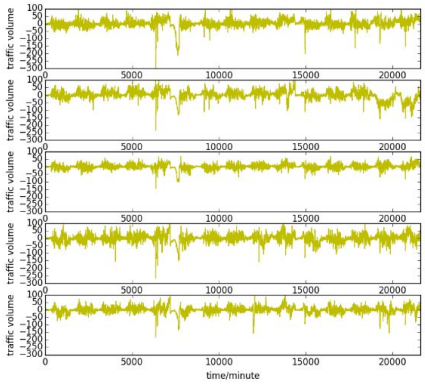

Fig. 6. 5 residual time series from the 3rd class (yellow color)

D. Origins of the clustering results

In Fig. 2, the nodes of some roads mainly belong to the same class. But in some roads, there are some nodes come from different class. This is caused by the influence of ramp in flows that belong to different classes.

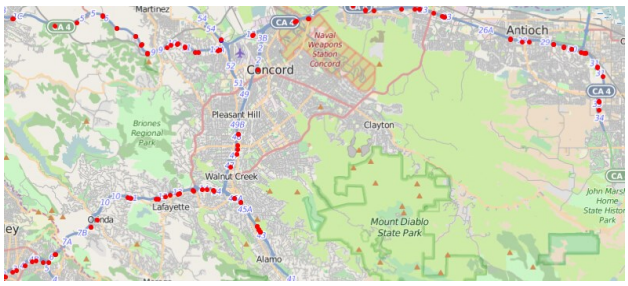

Fig. 7. A zoom-in study of the Concord area.

Fig. 7 shows that all nodes in the Concord are belong to the same class, since the road structure in this rural area is not complex and the fluctuation patterns are driven by same travel demand law.

Fig. 8. A zoom-in study of freeway US101 near Gilroy.

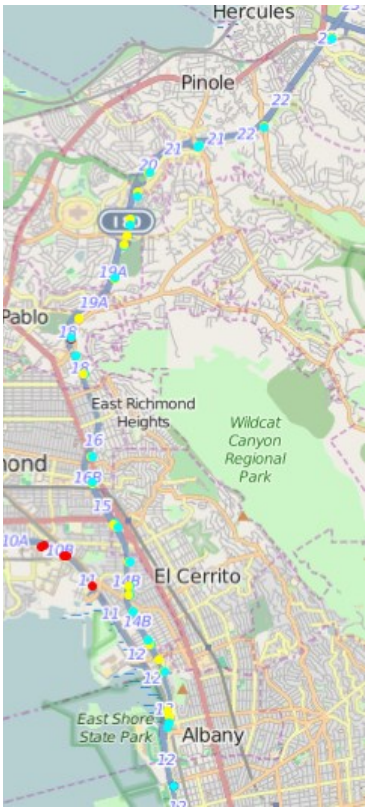

Fig. 9. Results of the US180

Fig. 8 shows the zoom-in study of freeway US101 near Gilroy. The flow direction of all blue nodes is south-to-north and the flow direction of all cyan nodes is north-to-south. The red nodes are in the freeway ramp entrance. We enlarge the map and observe in detail to get this conclusion.

Similar conclusions can be reached for freeway US180 near El Cerrito, except that the nodes in freeway ramp entrances may have different fluctuation patterns; see Fig.9. Check all the nodes, we can conclude the nodes in same road but different direction may have different fluctuation patterns, since the nodes in different directions are driven by different travel demands.

Fig. 10 shows that there exist two kinds of nodes in the urban areas. We cannot find the clear spatial patterns of these two kinds of nodes, since the travel demands in urban area are complicated.

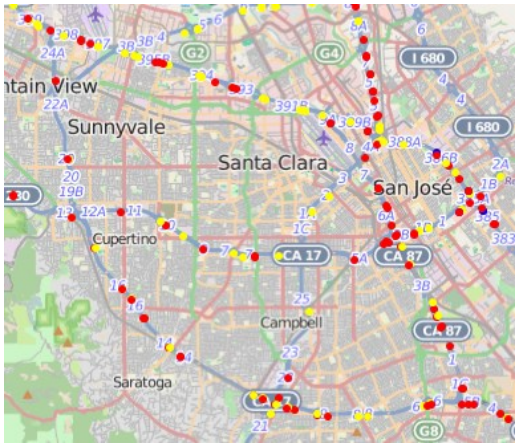

Fig. 10. Results of urban area

IV. Conclusion

In this paper, we discuss the clustering of the traffic time series that have similar fluctuation patterns. It needs to note that similar fluctuation pattern is not strictly equivalent to statistical correlation. We find that the clustering results are highly influenced by the travel demands patterns.

First, in the roads that are in rural area and are not directly linked to the urban area, the nodes mainly belong to the same class, since its fluctuation pattern is driven by same travel demand rule.

Second, in the roads that are in rural areas but are directly linked to the urban area, the nodes in the leaving and the entering directions may belong to two classes, since the travel demand laws of these two direction are often different.

Third, the nodes in the freeway ramp entrances may have different fluctuation patterns with the other nodes in the same road.

Fourth, the origins of the clustering results in urban area are complicated and we cannot find a simple law. We will study this interesting phenomenon in our future research.

The clustering of the traffic time series can be applied to many research fields. First, we can use this method to analyze the potential law of the traffic demand. Finding the law of traffic demand benefits to traffic management and traffic control. Second, the traffic time series of the same class have similar fluctuation pattern and these two time series may have correlation in a greater probability. When we build multi-sensor traffic prediction and missing data imputation model, we can first cluster the traffic time series, then select the potential correlation traffic time series within each class. This will reduce the computation times and will benefit to more accuracy prediction model. Moreover, this method can also be used to find the abnormal traffic time series. We will try to find more application of this method in our future research.

Acknoledgement

This work was supported in part by National Science and Technology Support Program 2013BAG18B00, Project Supported by Tsinghua University (20131089307), and Open Project of Beijing Key Laboratory of Urban Road Intelligent Traffic Control XN070.

References

[1] B. L. Smith, B. M. Williams, R. K. Oswald, "Comparison of parametric and nonparametric models for traffic flow forecasting," *Transportation Research Part C: Emerging Technologies*, vol. 10, no. 4, pp. 303-321, 2002.

[2] R. Chrobok, O. Kaumann, J. Wahle, M. Schreckenberg, "Different methods of traffic forecast based on real data," *European Journal of Operational Research*, vol. 155, no. 3, pp. 558-568, 2004.

[3] M. G. Karlaftis, E. I. Vlahogianni, "Statistical methods versus neural networks in transportation research: Differences, similarities and some insights," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 3, pp. 387-399, 2011.

[4] C. Chen, Y. Wang, L. Li, J. Hu, Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction," *Transportation Research Part C: Emerging Technologies*, vol. 22, pp. 103-118, 2012.

[5] E. I. Vlahogianni, M. G. Karlaftis, J. C. Golias, "Short-term traffic forecasting: Where we are and where we're going," *Transportation Research Part C: Emerging Technologies*, vol. 43, pp. 3-9, 2014.

[6] S. Sun, C. Zhang, G. Yu, N. Lu, F. Xiao, "Bayesian Network Methods for Traffic Flow Forecasting with Incomplete Data," *Lecture Notes in Computer Science*, Springer Berlin Heidelberg, vol. 3201, pp. 419-428, 2004.

[7] S. Sun, C. Zhang, G. Yu, "A bayesian network approach to traffic flow forecasting," *Intelligent Transportation Systems IEEE Transactions on* vol. 7, no. 1, pp. 124-132, 2006.

[8] L. Li, X. Su, Y. Wang, Y. Lin, Z. Li, Y. Li, "Robust causal dependence mining in big data network and its application to traffic flow predictions," *Transportation Research Part C: Emerging Technologies*, 2015.

[9] L. Qu, Y. Zhang, J. Hu, L. Jia, L. Li, "A BPCA based missing value imputing method for traffic flow volume data," *Proceedings of IEEE Intelligent Vehicle Symposium*, pp. 985-990, 2008.

[10] L. Qu, L. Li, Y. Zhang, J. Hu, "PPCA-Based missing data imputation for traffic flow volume: A systematical approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 10, no. 3, pp. 512-522, 2009.

[11] T. W. Liao, "Clustering of time series data-A survey," *Pattern Recognition*, vol. 38, no. 11, pp. 1857-1874, 2005.

[12] I. G. Guardiola, T. Leon, F. Mallor, "A functional approach to monitor and recognize patterns of daily traffic profiles," *Transportation. Research Part B: Methodological*, vol. 65, pp. 119-136, 2014.

[13] C. Chen, Y. Wang, L. Li, J. Hu, Z. Zhang, "The retrieval of intra-day trend and its influence on traffic prediction," *Transportation Research Part C: Emerging Technologies*, vol. 22, pp. 103-118, 2012.

[14] I.G. Guardiola, T. Leon, F. Mallor, "A functional approach to monitor and recognize patterns of daily traffic profiles," *Transportation Research Part B: Methodological*, vol. 65, pp. 119-136, 2014.

[15] G. W. Stewart, "On the early history of the singular value decomposition," *SIAM Review*, vol. 35, no. 4, pp. 551-566, 1993.

[16] I. T. Jolliffe, *Principal Component Analysis*, 2nd edition, New York: Springer-Verlag, 2002.

[17] A. Hyvärinen, E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks*, vol. 13, no. 4, pp. 411-430, 2000.

[18] Z. Zhao, Y. Zhang, J. Hu, L. Li, "Comparison study of PCA and ICA based traffic flow compression," *Journal of Highway and Transportation Research and Development (English Edition)*, vol. 4, no. 1, pp. 98-102, 2009.

[19] F. Douglis, J. Ousterhout, "Transparent process migration: Design alternatives and the Sprite implementation," *Software: Practice and Experience*, vol. 21, no. 8, pp. 757-785, 1991.

[20] M. Shindler, A. Wong, A. Meyerson, "Fast and accurate k-means for large datasets," *Advances in Neural Information Processing Systems*, pp. 2375-2383, 2011.

[21] R. A. Redner, H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195-239, 1984.

[22] G. McLachlan, T. Krishnan, *The EM Algorithm and Extensions*, John Wiley and Sons, 2007.

[23] PeMS, California Performance Measurement System. http://pems.eecs.berkeley.edu

[24] H. Zhou, Y. C. Soh, and X. Wu, "Integrated analysis of CFD data with K-means clustering algorithm and extreme learning machine for localized HVAC control," Applied Thermal Engineering, vol. 76, no. 5, pp. 98–104, 2015