Analysis of Network-level Traffic States using Locality Preservative Non-negative Matrix Factorization

Yufei HAN and Fabien Moutarde, *Member, IEEE*

***Abstract*—In this paper, we propose to perform clustering and temporal prediction on network-level traffic states of large-scale traffic networks. Rather than analyzing dynamics of traffic states on individual links, we study overall spatial configurations of traffic states in the whole network and temporal dynamics of global traffic states. With our analysis, we can not only find out typical spatial patterns of global traffic states in daily traffic scenes, but also acquire long-term general predictions of the spatial patterns, which could be used as prior knowledge for modeling temporal behaviors of traffic flows. For this purpose, we use a locality preservation constraints based non-negative matrix factorization (LPNMF) to obtain a low-dimensional representation of network-level traffic states. Clustering and temporal prediction are then performed on the proposed compact representation. Experiments on realistic simulated traffic data are provided to check and illustrate the validity of our proposed approach.**

I. Introduction

With developments of telecommunication, floating-car data, collected directly from vehicular mobile devices, become an essential and ever widely available data source for traffic data on large networks, including roads/streets for which no traffic monitoring infrastructure is available. Acquired floating-car data are employed to produce wide-coverage information on temporal properties of traffics flows, with which we can achieve global analysis of traffic patterns, and even predictions of traffic states several future time steps ahead. Through the processing of floating-car data, we are able to obtain helpful traveling information for vehicles, like estimated traveling time. Therefore, traffic data mining has become a hot research topic during recent years**.**

In previous research progress in traffic data mining, traditional methods use parametric models of traffic flows, in which a few parameters are calibrated with structural assumptions to simulate temporal evolution of traffic states [1]. In this kind of methods, cellular automata [2] is a typical instrument for powerful simulation and prediction systems. Data driven approaches, which adopt machine-learning techniques to extract statistical dependencies between data [3]-[5], become popular due to increasingly larger volume of collected floating-car data. These methods allow to "*let the data speak for itself*" and loosen assumed constraints of the proposed traffic dynamic model. Therefore, they are more flexible to describe and simulate temporal properties of traffic flows. However, in previous progress of both kinds of research, mining temporal patterns of traffic states measured on individual links plays a key role in traffic data analysis. Variations of traffic states in the whole network are described by analyzing temporal dynamics of traffic flows in each individual link. Actually, in a typical urban traffic scene, traffic states of one local region are highly correlated with neighboring areas. Such spatial configurations of traffic states can be used as prior knowledge during modeling traffic temporal dynamics for the whole network. They are useful in improving performances of traffic guidance.

In this paper, we propose to treat traffic states of all links in a large-scale link network as a whole, and to perform data mining task, clustering and long-term prediction on the network-level traffic states. Through our work, we aim to unveil typical spatial patterns and temporal dynamics of network-level traffic states, which provides overall descriptions traffic states over the whole link network. For large-scale urban traffic networks, network-level traffic information is often represented in a high-dimensional feature space, which makes it difficult to extract characteristics of global traffic states. In our work, we firstly adopt a geometrical weighted distance to evaluate similarity between network-level traffic patterns, which is described in section II.A. Then, we use a locality preservative non-negative matrix factorization method (LPNMF) to project network-level traffic state onto a compact representation model with much less dimensionality, as described in section II.B. In a further step, we perform clustering and temporal dynamic prediction on the low-dimensional LPNMF projection in section II.C. Finally in sections III and IV, we present clustering and prediction results of network-level traffic patterns on realistic simulated traffic data, and conclude the paper.

II. Geometrical Similarity Distance and Locality Preservative Non-negativity Factorization

A. Geometrical weighted similarity measure

Network-level traffic states are defined to be spatial configurations of link traffic states in a network, which is normally represented in *n*-dimensional vector, with *n* being the number of links in the network. Different network-level traffic states represent different global traffic state patterns. In a typical network, traffic state of one specific link is correlated with its up-stream or down-stream nearest neighbors in most cases.

Let links u_i^j and d_i^m respectively denote up-stream and down-stream nearest neighbors of link *i*. If link *i* is

Yufei Han and Fabien Moutarde are with Robotics Lab (CAOR) of Mines ParisTech, 60 Bd St Michel, 75006 Paris, FRANCE (e-mail: Yufei.Han@mines-paristech.fr, Fabien.Moutarde@mines-paristech.fr).

congested, its neighboring links u_i^j and d_i^m are more likely to be congested together than those located far from the link i and vice-versa. Motivated by this property, we adopt a weighted fusion among traffic states in geometrical neighborhoods to evaluate similarity between network-level traffic states. For the link i, we derive a weighted sum of the link-wise difference values with respect to the link i and its up-stream and down-stream neighbors, which is defined to be local variation v_i of traffic states around the link i, as expressed in Eq.1:

$$v_i = \sum_j w_j^u a(u_i^j) + \sum_m w_m^d a(d_i^m) + w^i a(i) \quad (1)$$

a is the link-wise difference between traffic states of the corresponding link. w_j^d, w_m^d and w^i are weights attached to up-stream neighbors, down-stream neighbors and the link i respectively. After that, we map L1 norm of $\{v(i)\}$ into [0,1] using a Gaussian kernel in Eq.2 as the final similarity measure between network-level traffic states:

$$S = \exp(-\frac{\sum_i v(i)}{2\delta^2}) \quad (2)$$

To normalize range of the weighted sum, the sum of all weights is required to be 1. The weight w^i corresponding to the link i should be the largest one. Without loss of generality, in this paper, we just treat that all neighboring share with the same weight value. By performing fusion among local neighborhoods, the derived similarity measure can be used as an indictor of spatial correlations between local neighborhoods.

B. LPNMF based network-level traffic state representation

Dimensionality of network-level traffic state representation is directly proportional to the number of links in the network. Given a large-scale network that is common with application, the resultant high-dimensional traffic state representation is difficult to store or use for analysis due to curse of dimensionality. To attack this issue, we propose to use locality preserving non-negative matrix factorization (LPNMF) [6][7] to obtain low-dimensional representation of global traffic states. Assuming that p samples of n-dimensional network-level traffic states are stored as $n*p$ matrix X, LPNMF factorize X into the non-negative $n*s$ matrix M and $s*p$ matrix V, which minimizes the following objective function:

$$O = \|X - MV\|_F^2 + \lambda Tr(VLV^T) \quad (3)$$

The first term is known as Frobenius reconstruction error. In this algorithm, each network-level traffic state is actually approximated by a linear combination of column vectors in M, weighted by components of the corresponding column vectors in V. Therefore, M can be regarded as a group of basis for representing global traffic states, while columns of V are s-dimensional coordinates of original traffic observations with respect to the basis M. In the setting of NMF, s is much less than the original dimensionality n. Therefore, V is a much lower dimensional representation of network-level traffic states after factorization. In contrast with SVD decomposition, derived manifold space is not necessarily orthogonal in NMF. Each data sample takes positive coordinates in the low-dimensional projection space. The above two properties makes NMF more suitable to describe the latent distribution structures, especially when overlap exists among different clusters of data samples. In the second term of the object function in Eq.3, L is Graph Laplacian [8], defined as $D - W$. In the matrix W, W_{ij} is a pair-wise geometrical weighted similarity measure matrix between the ith and jth network-level traffic state observation. D is a diagonal matrix whose entries are column sums of W, defined as Eq.4:

$$D_{ii} = \sum_j w_{ij} \quad (4)$$

By adding the Graph Laplacian based constraint, the obtained low-dimensional representation V are calibrated to keep similar topological structures as original data set X, which means that the similarity measure between the ith and jth column V_i and V_j reflects similarity of spatial patterns between the corresponding original network-level traffic state observations. Therefore, the low-dimensional LPNMF projection can denote structural information of general spatial configurations of traffic states in link networks, which makes it a suitable choice for performing predictive analysis of temporal dynamics.

C. Clustering and temporal prediction of network-level traffic states

According to the non-negativity property inherited from classical NMF settings, each component in V_i is proportional to contribution of the corresponding basis to represent general appearances of the original network-level traffic state observation X_i. Based on this property, we propose to use a simple scheme to determine cluster labels of each network-level traffic state observation. For each X_i, we examine V_i and assign X_i to the jth cluster if the jth component of V_i takes the largest value in V_i. Simply as it is, we can still find out the intrinsic distributional properties of network-level traffic states based on the LPNMF factorization method.

In our work, we introduce a k-Nearest-Neighbor (k-NN) based scheme to model temporal transition of network-level traffic states in a non-parametric way. Assuming we have a set of historic records of network-level traffic states $\{X_j^i\}$ (i=1…S, j= 1…T), which record traffic states for S different traffic scenes. Each scene contains T time sampling steps. We perform LPNMF on this data set. V_j^i corresponds to a LPNMF based representation of X_j^i. In a typical application of traffic state prediction, we usually have observed a sequence of network-level traffic states from the beginning time to the t-th time sampling step of one specific day and expect to predict how spatial configurations of network-level traffic states evolve in the following time until end of the day, which is a long term estimation of temporal dynamics of overall traffic states in the whole network. To solve this problem, we firstly project the sequence of currently obtained observations { $X_j^{'}$ } (j=1…t) onto low-dimensional representations {$V_j^{'}$} (j=1…t) with the basis M learned from historical data set. This procedure could be performed using Non-negative Least Squares (NNLS) [9], as illustrated in Eq.5.

$$V_j^{'} = \arg\min_{V_j^{'}} \left\| M V_j^{'} - X_j^{'} \right\| \quad V_j^{'} > 0 \qquad (5)$$

Due to fixed structures of the learned basis M and convexity of least square reconstruction, the obtained low-dimensional manifold representations {$V_j^{'}$} (j=1…t) have unique solutions. After that, we evaluate similarity between the obtained sequence {$V_j^{'}$} and the sub-sequences {V_j^{i}} (i=1…S, j=1..t) obtained from the historical records with the same time interval, following Eq.6 and 7:

$$sim_i = \exp\left(-\sum_{j=1}^{t} weight_j * d(V_j^{'}, V_j^{i})\right) \qquad (6)$$

$$weight_j = \exp(-a*(j-t)) \qquad (7)$$

$d$is cosine distance between the LPNMF based representations. According to Eq.6, the distance between temporal sequences of the low-dimensional representations is measured by a weighted sum of differences between the low-dimensional representations that are obtained at the corresponding time steps. The weight values are decayed exponentially as increasing interval values along the time axis between the stopping time t and a preceding sampling time j, which follows markovian assumption in time-series analysis [10]. Traffic states captured at earlier time than the current time t have less effecting on predictions of the future states. Based on the setting of distance measure, we can find the k nearest neighbor and their indices { $\boldsymbol{ind_m}$ } (m =1,2…,k) of the obtained sequence { $V_j^{'}$ } (j =1…t) in the historic records. Finally, predictions of unknown network-level traffic states for all following time steps from t+1 to T in the specific day are constructed using weighted average of the k nearest neighbors, as illustrated in Eq.8.

$$X_{t+i}^{'pred} = \sum_{p=1}^{K} \frac{sim_{ind_p}}{\sum_{p=1}^{K} sim_{ind_p}} X_{t+i}^{ind_p} \qquad (8)$$

Through k-NN operation, we aim to search for the first k traffic scenes in the link network that have with the most similar temporal evolution mode with the currently observed data. Due to fixed topological structures of the link network and regular patterns of demand and supply of traffic resources in daily traffic scenes, it is of high possibility that similar preceding temporal dynamics of traffic states leads to also similar future temporal behaviors of network-level traffic patterns. Therefore, as we can see in Eq.8, the contribution of each nearest neighbor is measured according to the similarity between historical records and the current observed data. Following this intuitive characteristic, k-NN based prediction can provide general descriptions of long-term dynamics of network-level traffic states with only one time of parsing in the historic records.

III. Experimental results of Clustering and temporal dynamics Prediction

A. Metropolis software and IAU-Paris Database

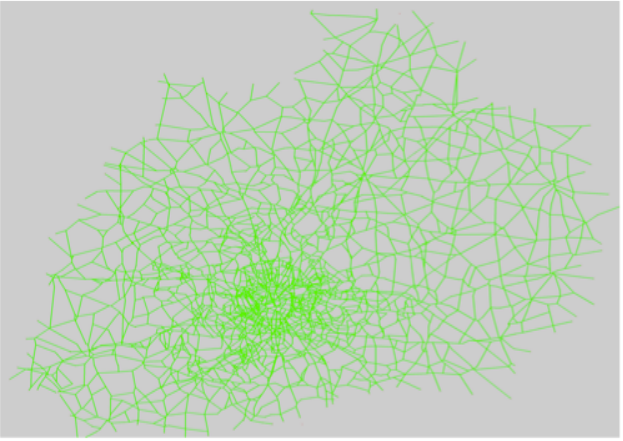


Fig. 1. Traffic network of Paris and suburb regions

To verify validity of the proposed method in clustering and modeling network-level traffic states, we firstly simulate real traffic scenes of the large-scale traffic network of Paris and its suburb regions using Metropolis [11], in order to generate a benchmark traffic database. Metropolis is a planning software that is designed to model transportation systems. It contains a complete environment to handle dynamic simulations of daily traffic in one specific traffic network, which allows the user to study impacts of transportation management policies in a large-scale urban traffic network in a time-dependent manner. The built traffic database is composed of 4660 intersections and 13627 links in the network shown on Fig. 1. Each simulated traffic scene is generated to cover 8 hours of traffic data observations, including congestion in morning rush hours. Different traffic situations are obtained by adding random events and fluctuation in the O-D matrix (Origin-Destination) and capacity of network flow. There are totally 108 simulated traffic scenarios in our benchmark data set. Each one contains 48 time steps, corresponding to 15-minute bins over which the network traffic flow are aggregated. To represent traffic states, we propose to use traffic index [12] in each link at a specific time, as in Eq.9.

$$x_{lt} = \frac{\Delta t_l^0}{\Delta t_{lt}} \qquad (9)$$

The denominator is the observed travel time in link l at time t, the nominator is the free-flow travel time on this link. The smaller the traffic index is, the corresponding link is more congested. To perform clustering analysis, we concatenate all observations of traffic states into a 13627*5184 matrix. Each column corresponds to a network-level traffic status obtained at each time step, which is a 13627-dimensional vector. In the experiment of clustering, the number of clusters is 3 and 5 respectively. For convenience of visualization, we project all the column vectors into 3-dimensional PCA space to illustrate structures of the obtained clusters. For modeling temporal dynamics of network-level traffic states, we set the number of LPNMF basis to be 30 to keep more information about spatial structures of global traffic states.

B. Clustering results of networked-level traffic states

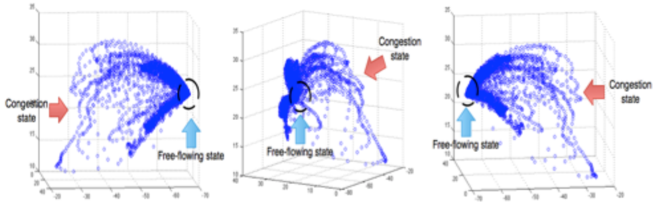


Fig. 2. Three-views diagram of network-level traffic states in 3D PCA space

We illustrate distributions of network-level traffic states in IAU-Paris database in 3D PCA space, as shown in three different viewpoints in Fig. 2. As we can see, the data points corresponding to the free-flowing network-level states concentrate within a small region in the PCA space. Compared with them, the data points corresponding to the scenes in which congestions occur in certain links are distributed rather sparsely and far from the region containing free-flowing states. Furthermore, while congestion in the link network become severer and severer, variations of data points become larger and larger. In fact, spatial configurations of network-level traffic states keep to be similar with each other if the whole network is almost free-flowing everywhere. On the contrary, congestion occurred at different parts of the network change spatial patterns of traffic states in different ways, which introduces large variations into distributions of network-level traffic patterns. We firstly divide all network-level traffic states in the database into three clusters, as we can see in Fig. 3.

The cluster labeled by blue legends represents that almost all links are free-flowing in the link network. Both red and dark green clusters indicate that traffic jam occurs in the certain parts of the link network. In Fig. 3, we select spatial configurations of network-level traffic states with the severest congestion in each cluster, which have the least average value of traffic indices among the corresponding clusters. They are used here as representative exemplars of spatial patterns of traffic states in each cluster. According to Fig. 3, in spatial configuration of each exemplar, red color is used to label congested links whose traffic indices are less than a specified threshold, while green color used for fluid links. We can see that the exemplar extracted from the red cluster contains much less busy links than the one from dark green cluster. It denotes that network-level traffic states in the dark green cluster contain severer congestion in the link network then the red cluster. Furthermore, as shown in both exemplars of the red and dark green clusters, most of congested links locate within the central region of the network. It implies that most traffic congestion occurs inside Paris. Suburb regions are free-flowing most of the time.

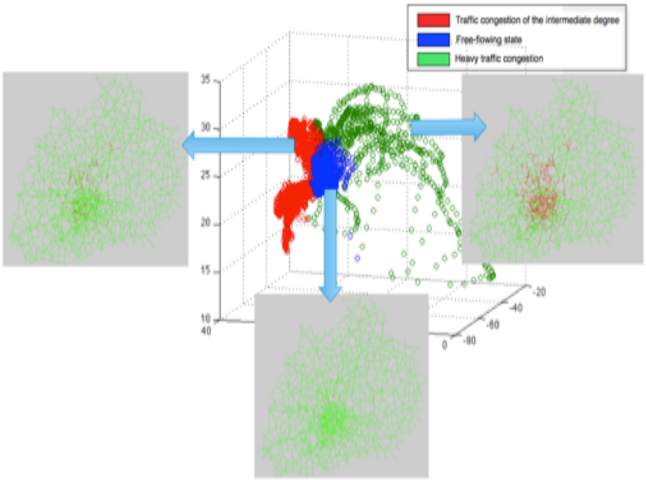


Fig. 3. Three clusters and exemplars of network-level traffic states

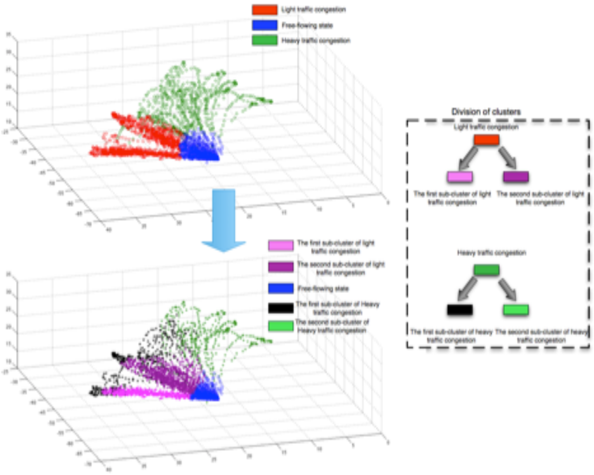


Fig. 4. Division of clusters after increasing the number of cluster

In Fig. 4, we increase the number of clusters from 3 to 5. Fig. 5 illustrates exemplars of clusters except the one corresponding to the free-flowing state, following the same settings in Fig. 3. We also compare structures of the three clusters shown in Fig. 3 and the five obtained clusters in Fig. 4. The cluster corresponding to network-level traffic state with light traffic congestion, labeled by red legends in Fig. 3, is further split into two parts that are labeled by pink and purple legends respectively in Fig. 4. These two sub-clusters have elongated shapes oriented to different directions in 3D-PCA space, which represents different distribution settings of congestion in the network. Exemplars of these two clusters illustrate the difference clearer, as shown in Fig. 5(a) and 5(b). In the exemplar of the sub-cluster labeled by pink legends, illustrated in Fig. 5(a), busy links tend to be closer to the central region than in the exemplar of the sub-cluster labeled by purple legends, as shown in Fig. 5(b). Despite of similar degrees of network-level congestion in both two exemplars, they indicate different spatial configurations of traffic states in the network, which is consistent with the difference of orientations of the two elongated sub-clusters.

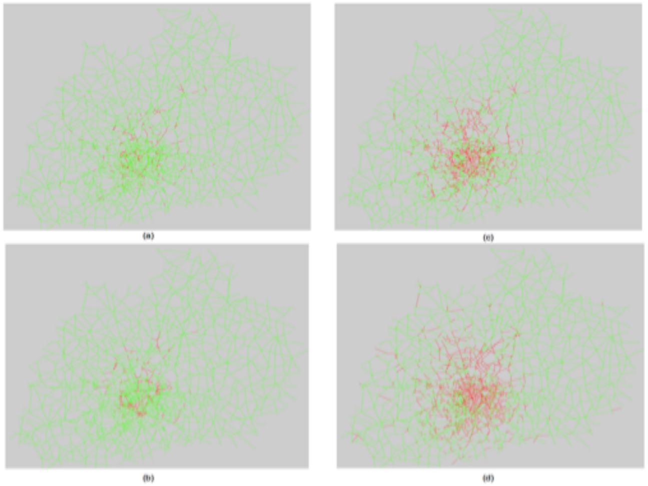


Fig. 5. Exemplars of sub-clusters. (a) and (b) are exemplars of sub-clusters labeled by pink and purple legends respectively. (c) and (d) are exemplars of sub-clusters labeled by black and light green legends respectively

Similar hierarchical way of splitting can also be observed in the dark green cluster in Fig. 3. As we can see in Fig. 4, this cluster is split to two sub-clusters labeled light green and black legends. Due to large variations of spatial configurations of traffic congestions, data points in both of two sub-clusters are sparsely distributed. However, these two sub-clusters still differ in spatial layout of network-level traffic congestion. In Fig. 5(c) and 5(d), we compare the exemplars of the two sub-clusters labeled by black and light green legends in Fig. 4 respectively. Generally, the exemplar in Fig. 5(d) contains more congested links. Furthermore, although the central region of the network is highly congested in both exemplars, the area to which network-level traffic congestion extend is more wide in the exemplar shown in Fig. 5(d), especially in suburb regions. This implies a different setting of traffic scenes during simulation.

C. Temporal prediction of network-level traffic states

We employ repeated random sub-sampling validation in our experiment: 88 of the whole 108 simulations of traffic scenes in the IAU-Paris database are selected randomly to form the historic observation records, and the remaining 20 are taken to be the testing set. Such random split is repeated for 200 times. In each split, in order to evaluate predicting accuracy between the estimated network-level traffic state $X_j^{'pred}(i)$ and the corresponding target $X_j^{'}(i)$ at the *j*th time step of the *i*th traffic scene, we calculate absolute difference s_i between mean of traffic index values in $X_j^{'pred}(i)$ and $X_j^{'}(i)$ as in Eq.10. Larger s_{ij} means less prediction accuracy.

$$s_{ij} = \left| mean(X_j^{'pred}(i)) - mean(X_j^{'}(i)) \right| \qquad (10)$$

For evaluating overall prediction performances for all time steps in the testing set, we calculate the average of all s_{ij} obtained in the testing data. Final overall evaluation result is then averaged over the total 200 iterations. In IAU-Paris database, there are 48 time steps of traffic observations in each simulation of traffic scenes. We choose the first 20 time steps as the observed sub-sequence of network-level traffic states, which covers early hours of each simulation. Long-term temporal dynamics of left 28 time steps are used to be targets of prediction. We compare the overall prediction performances of the proposed method with the historical data based prediction that uses average patterns of historical network-level traffic states at corresponding time steps as prediction results. The historical data based prediction is a baseline algorithm, because it doesn't make use of any heuristic knowledge about temporal dynamics of traffic states. Compared with it, k-NN operation in our method can select a group of historical traffic data that are more specific to the current traffic scenes. As a result, our method is expected to achieve higher accuracy in estimating spatial configurations of traffic states. In table 1 and Fig. 6, we compare overall prediction performances with different settings of the number k of the nearest neighbors.

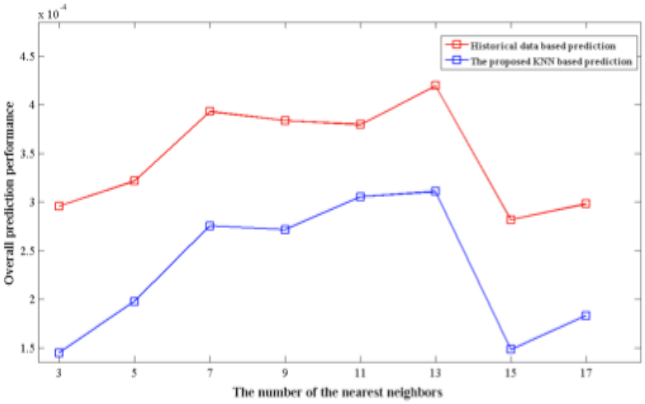
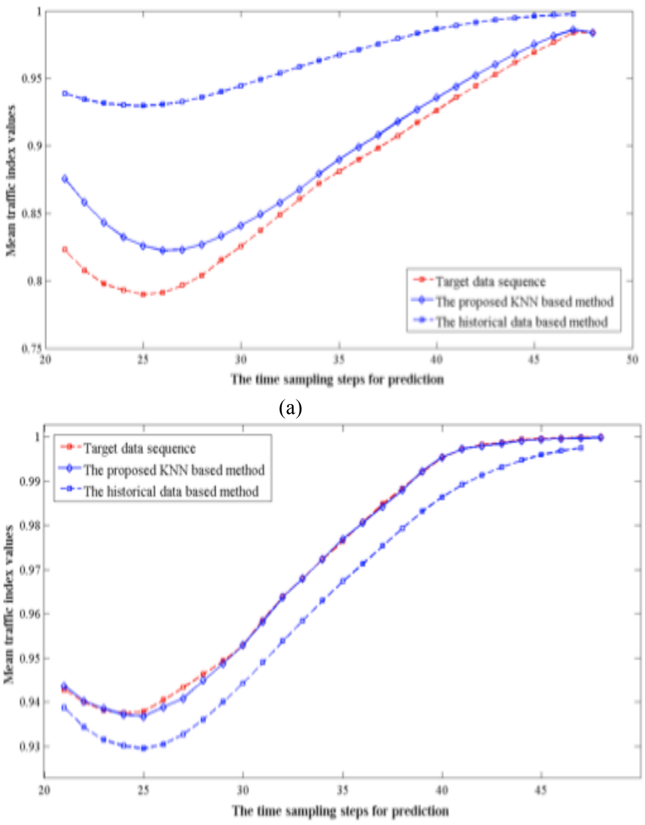


Fig. 6. Overall prediction performances with different settings of k in the proposed k-NN based method

TABLE 1. Prediction performances with different settings of k *in* k-*NN*

K	Historical data based method	KNN based method
3	2.96e-04	1.43e-04
5	3.22e-04	1.96e-04
7	3.92e-04	2.75e-04
9	3.84e-04	2.72e-04
11	3.80e-04	3.05e-04
13	4.20e-04	3.11e-04
15	2.82e-04	1.49e-04
17	2.99e-04	1.83e-04

According to table 1, we can find that average differences between predicted network-level traffic states and the ground truths are rather small. The main reason is that most links are free of traffic congestions in IAU-Paris database. Both of two methods involved in the comparison depend on historical records to reconstruct spatial configurations of network-level traffic states. Thus, prediction errors that are aroused in congestion regions of the link network become small. The variations of prediction accuracies with respect to both methods are caused by random selection of historical data set and testing set. Nevertheless, by comparing prediction performances of the k-NN based method with the historical data based one, it is obvious that the former achieves much better prediction than the latter, which confirms our idea that heuristic knowledge of temporal dynamic patterns is useful for long-term prediction in traffic data analysis. Furthermore, in Fig. 6, we can see that the difference of prediction performances between the two methods varies only a little by increasing k. It denotes that the top members in the nearest neighboring list play a dominant role in estimating the unknown temporal evolution patterns.



(a)

(b)

Fig. 7. (a) (b) Comparison of prediction performances on two different traffic scenes

Furthermore, Fig. 7 (a) and 7(b) show the subsequences of mean traffic indexes obtained from our proposed KNN based prediction, the historical data based prediction and ground truth in two traffic scenes of the testing data set in one iteration of random repeated random sub-sampling. The time sampling steps range from 21 to 48. Using the proposed KNN-based method, we can estimate temporal dynamic patterns more accurately than the historical data based method. Even around the turning point when network-level traffic states begin to recover from congestion, our method can still fit variation mode of traffic states in ground truth well, especially in Fig. 7(b). Interestingly, according to Fig. 7(a), the largest estimation error occurs during the time intervals around the turning point that contains more variations of spatial traffic state patterns than any other temporal periods.

IV. Conclusions

Our main contribution is to propose locality preservative non-negative matrix factorization (LP-NMF) to project high dimensional network-level traffic state observations into a smooth and compact manifold. Based on the derived low dimensional projection, we can describe typical spatial patterns and estimate long-term temporal dynamics of network-level traffic states more flexibly. Experimental results also indicate promising use of network-level traffic state modeling as prior knowledge in predicting temporal behaviors of global network traffic states.

Acknowledgment

This work was supported by grant ANR-08-SYSC-017 from the French National Research Agency. The authors specially thanks Cyril Furtlehner and Jean-Marc Lasgouttes for providing advice and the benchmark database used in this paper.

References

[1] A. Klar, R. Kuehne and R. Wegener "Mathematical models for vehicular traffic" in *Surv.Math.Ind*, vol. 6, p.215,1996.

[2] K. Nagel and M. Schreckenberg, "A celluar automaton model for freeway traffic", *Journal of Physics. I.2*, pp. 2221–2229,1992.

[3] B. Ghosch, B. Basu and M. O'Mahony, "Multivariate short-term traffic forecasting using time-series analysis", *IEEE Trans.Intell.Transport.Sys*.,vol.10,no.2,pp.246-254,2009.

[4] H. Kanoh et al., "Short-term traffic prediction using fuzzy c-means and cellular automata in a wide-area road network," in *Proceedings of the 8th International Conf. Intell. Transport. Sys*., 2005.

[5] W. Chun-Hsin, H. Jan-Ming and D. Lee "Travel-time prediction with support vector regression" *IEEE Trans. Intell. Transport. Syst.*, vol.5,no.4,pp.276,2004.

[6] D. Cai, X.F. He, X.Y. Wu, and J.W. Han, "Non-negative Matrix Factorization on Manifold", in *Proceedings of International Conf. Data Mining*, 2008.

[7] D. Cai, X. Fei He, X.H. Wang, H.J. Bao and J.W. Han, "Locality Preserving Nonnegative Matrix Factorization", In *Proceedings of International Joint Conf. Artificial Intelligence*, Pasadena, CA, 2009.

[8] F.R.K. Chung, "Spectral Graph Theory", in *Proceedings of AMS Regional Conference Series in Mathematics*, vol.92,1997.

[9] C.L. Lawson and R.J. Hanson, "Solving Least Squares Problems", *Prentice-Hall,* chapter 23, pp.161,1974.

[10] D.T. Crommelin and E. Vanden-Eijnden, "Fitting timeseries by continuous-time Markov chains: A quadratic programming approach", *Journal of Computational Physics*, vol. 217, pp. 782-805, 2006.

[11] F. Marchal, "Contribution to dynamic transportation models," *Ph.D.dissertation*, University of Cergy-Pontoise, 2001.

[12] C. Furtlehner, Y..Han, J.M. Lasgouttes, V. Martin, F. Marchal and F. Moutarde, "Spatial and Temporal Analysis of Traffic States on Large Scale Networks", In *Proceedings of Intelligent Transportation Systems Conference*, 2010.