Analysis of Large-Scale Traffic Dynamics in an Urban Transportation Network Using Non-Negative Tensor Factorization

Yufei Han, Fabien Moutarde

Analysis of Large-scale Traffic Dynamics in an Urban Transportation Network using Non-negative Tensor Factorization

Yufei HAN[*1] Fabien MOUTARDE[*2]
IMARA, INRIA Rocquencourt, France[*1]
(yfhan.hust@gmail.com)
Robotics Lab (CAOR), Mines ParisTech, PSL Research University, France[*2]
(60 Boulevard Saint-Michel, 75006 Paris, France, fabien.moutarde@mines-paristech.fr)

In this paper, we present our work on clustering and prediction of *temporal evolution* of global congestion configurations in a large-scale urban transportation network. Instead of looking into temporal variations of traffic flow states of individual links, we focus on temporal evolution of the complete spatial configuration of congestions over the network. In our work, we pursue to describe the typical temporal patterns of the global traffic states and achieve long-term prediction of the large-scale traffic evolution in a unified data-mining framework. To this end, we formulate this joint task using regularized Non-negative Tensor Factorization, which has been shown to be a useful analysis tool for spatio-temporal data sequences. Clustering and prediction are performed based on the compact tensor factorization results. The validity of the proposed spatio-temporal traffic data analysis method is shown on experiments using simulated realistic traffic data.

Keywords: *Large-scale traffic dynamics, non-negative tensor factorization*

1. Introduction

Recent decades witnessed rapid progress of distributed sensor systems in large-scale urban transportation networks [1][2]. Mobile GPS, loop magnetic detectors, etc, provide reliable ways to collect plentiful records of traffic flow states in transportation networks everyday [2][3][4][5][6][7]. Availability of the big traffic data stimulates great interests of researchers, engineers and public departments to extract underlined traffic dynamics, in order to understand traffic state evolution for better management of urban transportation. Along this direction, most published works of traffic data analysis focus on mining temporal dynamics of individual or small groups of links (either in arterial network or express ways) using model-driven [8][9][10][11][12][16][17][18] and data-driven methods [13][14][15][19][20][21][22][23][24]. The model-driven methods, like Cellular Automata [12] and other underlying physical models [8][9][10][11][16][17][18], are usually calibrated with structural assumptions to simulate temporal evolution of traffic states. Excellent as they are, the model-driven methods present less efficiency in modeling urban traffic. Congestion in urban transportation is easily subject to fluctuations induced by intersections, traffic signals, pedestrian behaviors and accidents. These fluctuations lead to complex spatio-temporal events, which makes it difficult to find a local stationary regime for the physical rule based traffic velocity models. In contrast, data-driven approaches describe spatio-temporal traffic flow patterns using statistical models based on historical observations. The latter are free from the structural constraints that are possibly biased from the real dynamics of traffic flows. Thus, they are more flexible to adjust and fit the underlined traffic state variations. In this branch, Kalman filter [14] and ARMA (Autoregressive Moving Average) [15], originated from state space theory, are popularly used to predict linear variation tendency of traffic flows [14][15][19]. In [20][21][22], neural networks [20][21] and hybrid non-linear dynamic systems [22] are used to approximate short-term non-linear fluctuations of traffic flow states. Due to the intrinsic multiple-input and multiple-output (MIMO) structures, neural networks intrinsically integrate spatio-temporal correlations between local link segments. In [23][24], spatio-temporal correlations between local links are also modeled using Markov Random Field [25] and Multi-Agent System [26]. These inspiring works concatenate global structural information of transportation networks to improve descriptive power of traffic flow models. However, when one tries to apply data-driven methods to large-scale network, high dimensionality of spatial network structure results in curse-of-dimensionality in model configuration. An interesting way of handling complexity and dimensionality issues is to partition the full network in sub-regions, with the congestion behavior of each area modeled by Macroscopic Fundamental Diagram (MFD). MFD is firstly defined by Geroliminis et al in [27]. It provides an approximate relation between space-mean flow, vehicle speed and vehicle density of sub-region, which serves as a macroscopic measure of its capacity. The first thorough empirical investigation of MFD for analyzing urban traffic can be found in [28].

Spatial partitioning of network based on MFD, as proposed in [29], is quite promising for designing traffic management strategies at city level. However, our goal is to unveil typical temporal congestion daily evolutions, rather than provide a geographical partition.

Classical model-driven or data-driven methods are proposed to estimate temporal traffic state dynamics of individual links, while ignoring spatio-temporal correlation between links and successive time intervals. It causes the prohibitive model complexity when they are applied directly to analyze traffic states of a large-scale network. To attack this issue, we propose to investigate temporal evolution of spatial co-occurrence patterns of congestion level in the entire network during specific time intervals. Such macroscopic view of traffic dynamics can identify bottleneck of transportation networks in different time slots and improve traffic management strategies dynamically. Drivers can use global traffic state information to optimize their traveling. The analysis result can be also used as a spatial consistency constraint of traffic congestion level in modeling the temporal variations of traffic states of individual links.

Our previous work [30] used matrix factorization to derive low-dimensional representation of the global traffic state and identify typical spatial congestion co-occurrence patterns of the whole network. Despite of the achievement, this work is not designed intrinsically to find out the temporal causality of congestion during successive time intervals. In this paper, we follow principles of matrix factorization and use regularized non-negative *tensor* factorization (r-NTF) to extract a compact representation of this temporal evolution of the spatial congestion pattern. K-means clustering [31] is then performed on the derived representation to unveil typical temporal evolution of the spatial congestion patterns. Using the proposed r-NTF, we further achieve long-term temporal prediction of spatial congestion patterns in the network.

Section 2 introduces the proposed r-NTF method employed in the analysis. Section 3 presents the simulated traffic data of a large-scale urban network, used as data source in the following analysis. In Section 4, we illustrate detailed clustering results of typical spatio-temporal congestion patterns obtained through the NTF projection. Section 5 further performs a long-term prediction on temporal behaviors of spatial congestion configurations based on the NTF scheme. Section 6 concludes the work.

2. Spatio-temporal analysis of traffic data using Non-negative Tensor Factorization

2.1. Basics about tensor

A tensor T is formally defined as a multi-way array $R^{N_1 \times N_2 \times \ldots \times N_M}$, where M is the order of tensor (i.e. its number of ways). N_i represents the dimensionality of the i-th order of the tensor. Through this way, a single entry $T_{i_1, i_2 \ldots i_M}$ in the tensor is localized by fixing the total M indices $\{i_1, i_2 \ldots, i_M\}$. A review of tensor is provided in Kolda et al. [32][33]. The multi-way structure of tensor provides a natural way to encode the underlying multiple dependencies in the sequential multivariate data. For example, in video processing, tensors are widely used to represent temporal streams of multi-dimensional data, such as 2D images in video frames [34] and user product rating profiles in recommendation systems [35].

2.2. Representing spatio-temporal traffic state data using multi-way tensor

The traffic state data are composed of congestion levels for *all* links in a large-scale urban transportation network. For each link, temporal variation of congestion is described by sampling each period (e.g. one day) with a fixed and shorter interval over which congestion level is estimated and averaged. Our purpose is then to investigate spatio-temporal dynamic patterns of traffic congestion in the network during the concerned period. To this end, we use a tensor $T \in R^{n \times m \times l}$ to represent temporal sequences of traffic state measurements. Integers `n`, `m` and `l` respectively correspond to the number of links in the network, the number of sampling steps in each time period, and the number of concerned periods (typically half-days or days) within total historic observations. Each entry $T_{i,j,k}$ represents the traffic flow state of the `i`-th link observed at the `j`-th sampling step of the `k`-th time period. Each column vector $T_{:,j,k}$ of the tensor concatenates traffic states of all links at a specific sampling step into one `n`-dimensional vector. It encodes spatial congestion co-occurrence over entire network. We note it as network-level traffic state, inherited from our previous work [29]. Each frontal slice $T_{:,:,k}$ stores the network-level traffic states within the same period `k` as its column vectors following their temporal orders. It represents temporal variations of network-level traffic states $T_{:,j,k}$ within the concerned period `k`. We note it using spatio-temporal traffic state dynamics in this paper. It describes temporal dynamics of spatial congestion co-occurrence in the entire network. Tensor factorization is then used to derive compact representation of the spatio-temporal traffic state dynamics of each time period.

2.3. Regularized Non-negative Tensor Factorization for spatio-temporal traffic data analysis

We propose Regularized Non-negative Tensor Factorization (r-NTF) for analyzing spatio-temporal

traffic dynamics. The basic idea is to decompose one tensor as a canonical combination of low-order structures, such as vectors and matrices, with an additional regularization term to guarantee smoothness of factorization results. Components in the canonical combination are factors that contribute separately in reconstructing the tensor, corresponding to physical sense of each way of the tensor. For better understanding, we firstly review Non-negative Matrix Factorization (NMF) used in our previous work [29] as an example. NMF [36][37] is as a special case of r-NTF applied in matrices (2nd order tensors) as illustrated in Eq.1:

$$\arg\min_{U\geq 0, V\geq 0} \left|\left| M - \sum_{i=1}^{r} (U_{:,i} \circ V_{:,i}) \right|\right|_F \quad (1)$$

where M is a $n \times m$ matrix, $U_{:,i}$ and $V_{:,i}$ are the `i`-th columns of the non-negative matrices U $\in R^{n\times r}$ and V $\in R^{m\times r}$. Operator ∘ is the outer product of vectors. $||\ ||_F$ is Frobenius norm of matrix. As presented in [29], successive columns of M correspond to a temporal sequence of network-level traffic states collected at different temporal sampling steps, each column of V is the NMF projection of the corresponding network-level traffic states. Integer `r` is the dimensionality of the NMF projection space spanned by columns of U. Normally, `r` is set to be much less than the row dimension of M, thus the column space of V forms a low-dimensional representation of network-level traffic states. The non-negativity constraint on U and V makes each network-level traffic state $M_{:,j} \in R^n$ approximated by an additive linear superposition of the column space of U of NMF [29] as in Eq.2:

$$M_{:,j} = \sum_{i=1}^{r} U_{:,i} V_{i,j} \quad (2)$$

where $M_{:,j}$ and $U_{:,i}$ are the `j`-th column of M and the `i`-th column of U respectively. $V_{i,j}$ is the element located at the `j`-th column and `i`-th row of V. Columns of U are the learned bases representing typical spatially structural patterns of congestion. $V_{i,j}$ represents to which degree the `j`-th network-level traffic state $M_{:,j}$ is associated with the learned basis vector $U_{:,i}$. Thus, through the linear projection, NMF generates a low-dimensional signature depicting spatial characteristics of network-level traffic states. However, due to the two-way structure of matrix, NMF has limited resolution in temporal analysis. The ranging order of columns $M_{:,j}$ mixes temporal sampling steps of different periods, thus NMF cannot distinguish spatio-temporal traffic evolutions of different time periods directly.

Therefore, preserving the descriptive power of NMF while extending the methodology to investigate spatio-temporal traffic dynamics, we propose the r-NTF as follows:

$$\arg\min_{U\geq 0, V\geq 0, Q\geq 0} \left|\left| T - \sum_{i=1}^{r} (U_{:,i} \circ V_{:,i} \circ Q_{:,i}) \right|\right|_{L^2} + \lambda Tr(Q^T L Q) \quad (3)$$

$$L = D - W \quad (4)$$

$$D_{i,i} = \sum_j W_{i,j} \quad (5)$$

where T is the 3-way tensor storing the traffic data as described in Section 2.1. Vectors $U_{:,i}$, $V_{:,i}$ and $Q_{:,i}$ are `i`-th columns of three non-negative matrices: U of size `n×r`, V of size `m×r`, and Q of size `l×r`. $||\ ||_{L^2}$ denotes the entry-wise sum of square. Operator Tr is matrix trace and λ is the regularization parameter. Similar to NMF, the column spaces of U, V and Q form the `r`-dimensional representations of respectively link-specific, temporal sampling step specific, and time period specific factors in spatio-temporal dynamics of traffic states. According to Eq.3, by fixing U and V, we approximate each frontal slice of the 3-way tensor $T_{:,:,k}$ as a strictly additional basis expansion shown in Eq.6:

$$T_{:,:,k} = \sum_{i=1}^{r} Q_{i,k} (U_{:,i} V_{:,i}^T) \quad (6)$$

$Q_{i,k}$ is the entry locating at the `k`-th column and `i`-th row of Q. Matrix $U_{:,i} V_{:,i}^T$ has the same size as the frontal slice $T_{:,:,k}$. Considering $U_{:,i} V_{:,i}^T$ (i=1,2,…,r) as a set of matrix-form bases, Eq.3 illustrates a linear projection of the frontal slice to the space spanned by the bases. $\{Q_{i,k}\}(i = 1,2...,r)$ are the `r`-dimensional projection coordinates of $T_{:,:,k}$ with respect to the bases. Note that we borrow the notion of multivariate signal expansion and use them for the matrix-form objects. That provides us an intuitive understanding of tensor factorization. As in NMF, benefited from the non-negativity constraint, the linear expansion in Eq.3 is an additional superposition of the bases $\{U_{:,i} V_{:,i}^T\}$ approximating $T_{:,:,k}$. Thus, each matrix-form basis $U_{:,i} V_{:,i}^T$ implies an underlying component of the frontal slice. In our case, since each frontal slice $T_{:,:,k}$ records temporal variation of network-level traffic states within each time period, the matrix-form bases $\{U_{:,i} V_{:,i}^T\}$ correspond to components forming the large-scale traffic temporal dynamics, a.k.a. part-based representation of $T_{:,:,k}$. Given this property, the column space of the projection $\{Q_{i,k}\}(i = 1,2...,r)$ forms `r`-dimensional signatures of spatio-temporal traffic state pattern of each concerned time period in $T_{:,:,k}$.

The second term in Eq.3 is the structural

regularization of the factorization. In this term, L is called Graph Laplacian [38][39] as defined in Eq.4. In the matrix W, the element $W_{i,j}$ located at i-th row and j-th column, is the pair-wise similarity measure between spatio-temporal traffic state dynamics of the i-th and j-th time period, corresponding to the i-th and j-th frontal slice of T. This similarity measure is either obtained according to empirical analysis on historic observations, or defined based on knowledge of experts in the transport domain. According to Eq.5, D is a diagonal matrix whose entries are column sums of W. Graph Laplacian originates from spectral graph theory [39][40]. By adding the Graph Laplacian based constraints, the obtained low-dimensional representation V is calibrated to have similar geometrical structures as the original data T without increasing further computation cost. In Eq.5, the similarity measure W encodes our prior knowledge about the underlying spatio-temporal traffic dynamics. The Graph Laplacian constraint in Eq.4 thus requires the derived tensor projection of the traffic data to be consistent with intrinsic data distribution information. Besides, it also makes the projection smooth enough to be compatible with data analysis algorithms, such as clustering and temporal prediction models.

2.4. Clustering and long-term temporal prediction of large-scale traffic dynamics

Through tensor factorization, each column vector $Q_{:,k}$ of Q represents a r-dimensional signature of spatio-temporal traffic evolution of the corresponding time period. We propose to perform K-means clustering algorithm on the column space of Q. The centers of the clusters correspond to typical spatio-temporal traffic dynamics patterns. Without tensor factorization, the only way to achieve this goal is to perform clustering directly on the frontal slices of the tensor. Due to high-dimensionality of the frontal slices, it is difficult to find distinct cluster structures in Euclidian space of frontal slices [35]. In contrast, the r-dimensional signatures derived from the tensor factorization form a projection subspace with much lower dimensionality, thus strengthen the underlying cluster structure of spatio-temporal traffic dynamic patterns. As indicated in [41], K-means does not necessarily guarantee to find the most optimal assignment configuration, corresponding to the global objective function minimum. The algorithm is also significantly sensitive to the initial selected cluster centers. We tackle this issue by running K-means for multiple times with randomly initialized clustering configuration at each time. Finally we choose the clustering result that minimizes most the sum-of-square k-means objective function.

Based on the clustering analysis of large-scale spatio-temporal traffic dynamics, we further aim to achieve long-term temporal prediction of the spatial congestion co-occurrence in the entire network with the proposed tensor factorization framework. This work follows a simple principal: for two temporal sequences of network-level traffic states corresponding to two different periods, given similar tendency of spatio-temporal traffic dynamics at the beginning sampling steps of the concerned time period, they are likely to have similar forthcoming temporal congestion evolution during the rest time of the period. In urban transportation networks, the topological structure of the network is stable and drivers' behaviors are historically consistent. Plenty of historic data thus provide comprehensive information about typical spatio-temporal traffic dynamics within the fixed time period, such as consistent occurrence of morning and evening peaking hours during working days. As a result, we can achieve temporal prediction of network-level traffic state evolution by heuristically comparing the partially observed traffic dynamics and large-scale traffic dynamic patterns in historic data.

To formulate this task using the proposed tensor factorization framework, we assume all historic observations of traffic states stored in a tensor structure $T^{historic} \in R^{n \times m \times l}$ with l the total number of time periods involved in historic observations. Given another specific time period, we suppose observed only the first m_1 time steps of network-level traffic states. The temporal prediction then estimates spatio-temporal traffic evolution from the ($m_1 + 1$)-th step until the end. The central idea of this work is two-folds. Firstly, compared with one-step ahead forecast, the prediction task aims to cover a dynamical process of much longer time period. We name it as "long-term prediction" in contrast with short-term estimation, such as forecast of 5-10 minutes ahead. Furthermore, different from previous work focusing on estimation of precise traffic state (traveling time or speed) of one single link, we aim to estimate temporal evolution of spatial congestion co-occurrences of the whole network. It provides an overall description of spatio-temporal dynamic patterns of congestions. For management departments of large-scale transportation network, the macroscopic description of traffic dynamics in both spatial and temporal scale is helpful to identify bottlenecks of the network in different time slots, in order to globally optimize use of traffic resources.

To achieve the long-term temporal prediction, we firstly perform r-NTF on $T^{historic}$ to project the historical spatio-temporal traffic dynamics of the fixed time period to r-dimensional subspace, as seen in Eq.7:

$$T^{historic} \approx \sum_{i=1}^{r} \left(U_i^{historic} \circ V_i^{historic} \circ Q_i^{historic} \right) \qquad (7)$$

Traffic dynamics of the specific partially observed period is represented as a frontal slice M. It can be considered as a point lying in the r-dimensional smooth

subspace spanned by the learned base $\{U_i^{historic} \circ V_i^{historic}\}$. Reconstruction of the missing entries in the frontal slice M is formalized as follows:

$$q = \arg\min_{q \geq 0} \left\| M - \sum_{i=1}^{r} q_i (U_i^{historic} \circ V_i^{historic}) \right\|_{L^2}^2 + \lambda_M \sum_{j=1}^{K} s_{h_j} \left\| q - Q_{:,h_j}^{historic} \right\|_{L^2}^2 \quad (8)$$

$$\tilde{M} = \sum_{i=1}^{r} q_i (U_{:,j}^{historic} \circ V_{:,j}^{historic}) \quad (9)$$

where q is the estimated r-dimensional projection coordinates of M with respect to the learned bases $\{U_i^{historic} \circ V_i^{historic}\}$. We select K-nearest neighbors of the partially-observed time period M among all time periods in the historical data, corresponding to the frontal slices { $T_{:,:,i}^{historic}$ } (i=1,2,…,l). The K-nearest neighboring frontal slice is indexed by $\{h_j\}(j = 1, 2, ..., K)$. s_{h_j} is the Euclidean distance based similarity measure evaluated between M and the selected frontal slice $T_{:,:,h_j}^{historic}$. $q_{h_j}^{historic}$ are the r-dimensional signatures of the K-nearest neighboring frontal slices in the historic records derived from the tensor factorization. The first term in Eq.8 aims to minimize the reconstruction error between the observed traffic states and their corresponding estimates in $\tilde{M}$, forcing $\tilde{M}$ better approximate the observed information. The second term performs the Graph Laplacian constraint on the signature q of $\tilde{M}$. By minimizing this regularization term, we aim to find a projection of the partially observed M that optimally approximate the observed information and preserve the topological relations with the projections of the nearest neighboring frontal slice in the historic data. The basic idea behind the design is the principle of temporal causality in traffic dynamics that we employ to achieve the long-term temporal prediction. The solution of Eq.8 is obtained by iteratively updating q using KKT condition [42]. Each step of the update is formulated in Eq.10.

$$q_i^{new} = \frac{q_i^{old} (XYV^T + \lambda_M \sum_{j=1}^{K} s_{h_j} q_{h_j}^{historic})_i}{(q^{old} VYV^T + \lambda_M q^{old} \sum_{j=1}^{K} s_{h_j})_i} \quad (10)$$

q_i^{new} and q_i^{old} are the i-th entry of q after and before the update. Y is a $n \times m$ dimensional vector that arranges all elements of M into a vector following a column-wise order. R is a vector of the same size as X, which is a binary mask that gives 1 to the observed entries in X and 0 to the missing ones. $V \in R^{(n \times m) \times r}$ is a matrix with each row as the Kronecker product of $U_{:,j}^{historic}$ and $V_{:,j}^{historic}$. Once the optimum estimate of q is obtained, the missing entries of M, corresponding to the unobserved future traffic states, are reconstructed using Eq.9, and used as the prediction.

3. IAURIF Traffic Database

3.1. Metropolis traffic simulation software

The benchmark IAURIF database used to verify the validity of the proposed r-NTF based method is generated by simulating real-traffic sequences of a large-scale traffic network using Metropolis [43][44]. Metropolis is a planning software designed to model urban transportation systems. It allows the user to study impacts of transportation management policies for metropolitan areas and their fringes in a time-dependent framework. Metropolis simulates commuters' traveling behaviors and congestion in urban areas. The core of the simulation system is a dynamic simulator that integrates joint commuters' departure time and their choices of routes in the transportation network. During simulation of traffic sequences, each commuter is characterized by specific parameter values individually [43][44]. At any moment, locations of all commuters are known. Given the geometrical structure of network, and traffic demand specified by an Origin-Destination (O-D) matrix, commuters choose the fastest path from their origin to destination. Interaction between individual route choices is taken into account by running several iterations of the simulation until stabilization of all choices of itineraries and departure times. Metropolis does not simulate in details the physics of traffic, but rather estimates the travel time on a link as a function of density of cars and incoming flow. Each link is therefore modeled as a queue in which the car remains during estimated travel-time. By default, the linear speed-density fundamental diagram is used: each link is characterized by a free-flow velocity V_{ff} (typically related to speed-limit) and a traffic-freezing maximum density ρ_{max}; current average speed on the link is estimated as $V=V_{ff} \times(\rho_{max}-\rho)/\rho_{max}$, therefore decreasing from V_{ff} to zero when density on the link increases from zero to ρ_{max}. This traffic model is relatively simplistic, firstly due to assumption of linear speed-density law, secondly because it supposes that all vehicles present on the link at a given time shall experience same travel-time in the link; also traffic lights along the link are not explicitly taken into account. However, in our present work, what we need is only simulations that are realistic enough to provide a good testing benchmark for our traffic evolution analysis, clustering and prediction framework. An interesting feature of Metropolis from this point of view is that by modulating amplitude and geometry of the O-D matrix, we can easily generate varying spatio-temporal congestion patterns.

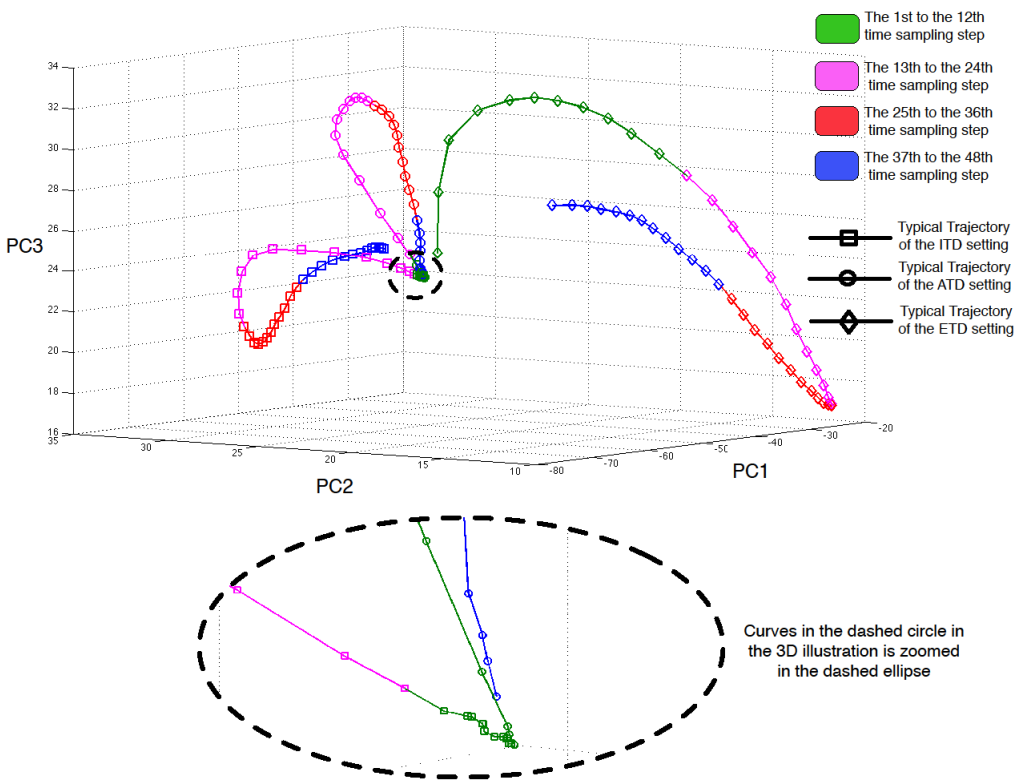
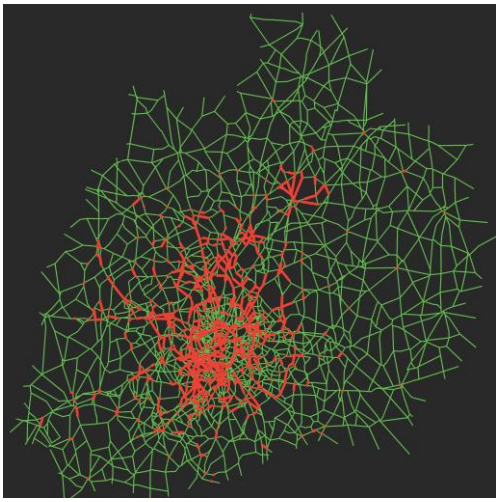


Figure 1. Visualization, in 3D PCA space, of three typical trajectories corresponding each to one half-day simulation for one of the three settings of traffic demands

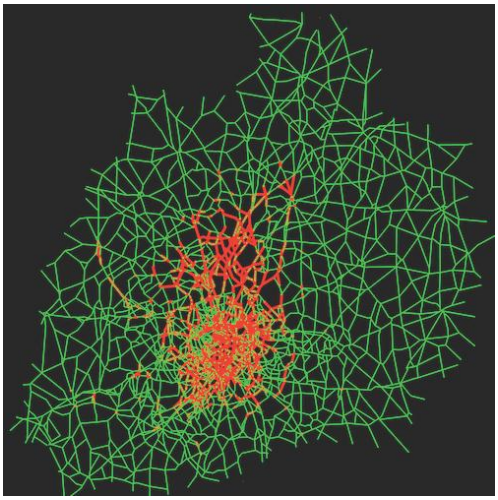
3.2. Settings of IAURIF database

The network used in IAURIF database contains totally 13627 links of Paris and its suburb region, as displayed in Figure 2. There are totally 146 simulated traffic sequences in the data set. Each simulated traffic sequence covers 8 hours of traffic data observations, involving congestion in peak time. Total 48 time sampling steps within each simulation divide the whole 8 hours into 10-minute bins over which the network traffic flows are aggregated. The typical temporal evolutions are shown on Figure 1, plotted in 3D PCA space. To represent local traffic state of each link, we use traffic index [43][44] defined as ratio between the minimum traveling time of vehicles on this link given the upper limit of speed and the observed average traveling time on the link during each time interval. The traffic index belongs to]0,1]: if the index value equals to 1, it means the vehicles run at their full speed on the link, indicating the link is free-flowing; as the traffic index decreases, the corresponding link becomes more congested. We store all traffic index values into a three-way tensor $T \in R^{13627\times 48\times 146}$ following descriptions in Section 2.

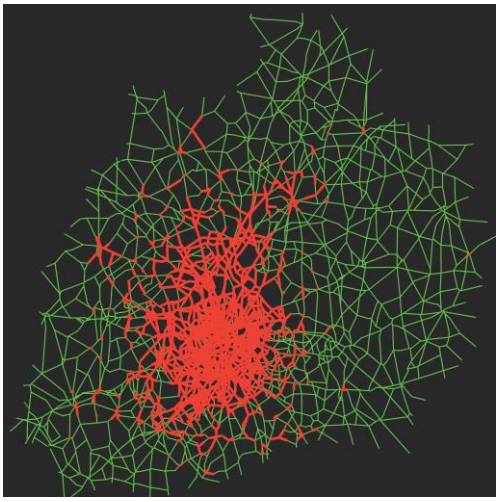
The simulated traffic sequences involve three different configurations of O-D matrix of all 3000 commuters in the network. They start their travel from outskirts of Paris into the central area within 8 hours in each simulation. With this setting, we aim to describe traffic behaviors during the morning peak hour of the Paris transportation network. Different configurations of O-D matrix result in variations of global congestion level and different spatial distribution of congestion during peak hour. In the first setting, traffic demands are distributed relatively evenly in the outskirt area near the central Paris, leading to light isotropic congestion inside and around the central Paris. For the second case, we set more travel plans from the northern outskirt area to the central Paris, which produces local congestion patterns in both the northern outskirt and central region. Furthermore, we add random variance to the total amount of travel plans contained in the O-D matrix, covering both globally light and heavy congestion that share the same specific spatial congestion patterns in the network. In the third case, we increase travel paths inside the central and northern area to cause extremely heavy traffic burden in the corresponding areas. As a result, the derived sequences of traffic flow states present global congestion ever since the beginning of simulation. They are used to simulate occurrence of unexpected extreme incidents in the network. We name the three settings as “Isotropic Traffic Demand” (ITD), “Anisotropic Traffic Demand” (ATD) and “Extreme Traffic Demand” (ETD).



(a)



(b)



(c)

Figure 2. (a) typical spatial congestion patterns of the ITD setting; (b) typical spatial congestion patterns of the ATD setting; (c) typical spatial congestion patterns of the ETD setting

To understand the traffic demand settings intuitively, Figures 2(a) and 2(b) illustrates the typical spatial congestion patterns in the network corresponding to the first two traffic demand settings during the peaking hour (from the 15th to the 25th time sampling steps). For the third setting, Figure 2(c) shows typical congestion patterns corresponding during the peak hours, indicating the occurrence of utterly heavy congestion over the network. To visualize the distribution of congestion in the network, we use a fixed threshold to identify the occurrence of congestion in each link. Links with traffic index less than 0.4 are illustrated as congested links, indicated by red legends. Otherwise, they are shown as free-flowing links, labeled by green legends. The threshold is chosen according to the quintile of all traffic indexes in the simulation.

Following the three settings, we generate 37, 91, 18 simulations respectively. In the ATD setting, due to the inserted random variances of traffic demands, we generate 39 simulations among all 91 simulations containing light global congestion level. The rest 52 correspond to the occurrence heavy global congestion level. To visualize distribution of the network-level traffic states and its temporal evolution, we project all 13627 dimensional network-level traffic state vector into 3-dimensional PCA space [45]. In the figures illustrating data distribution and clustering results in the PCA projection space, saying Figures 1, 4 and 5, the three axes of the plot correspond to the top three principal components in PCA projections that keep most variances of the original network-level traffic states. The labels ‘PC1’, ‘PC2’ and ‘PC3’ are used to denote the first, the second and the third principle components respectively in the figures. The principal components are derived from eigen-decomposition of the empirical covariance matrix of network-level traffic states. Thus they are dimensionless quantities and have unit of 1. The observations of the global free flowing states are distributed within a small region compactly. In contrast, those of medium or severe congestion are distributed sparsely and biased from the region of the free flowing state. Spatial configurations of local traffic states keep the same if the whole network is global free flowing. On the contrary, congestion occurred at different parts of the network changes the spatial configurations in different ways, increasing variations of global traffic patterns. In Figure 1, we link network-level traffic states of the same simulated sequences following their temporal orders in the PCA projection space. The resultant trajectory represents temporal evolution of the network-level traffic states, producing an intuitive view of the spatio-temporal traffic state dynamics. Different markers on trajectories are used to indicate different traffic demand settings. In each trajectory, color legends are used to indicate successive time intervals. The typical trajectories of the ITD and ATD setting have different orientations in the PCA space, consistent with difference of spatial patterns

of congestion. All trajectories start from global free flowing state, as we initialize all simulations with global free flowing state. Trajectories of the ITD and ATD settings converge to the free flowing state, indicating the network restores its fluidity after peak hour. In contrast, the ETD setting leads to very severe congestion in the network, with some links congested even at the end of simulations. Thus the corresponding trajectory is different from the other two, and converges to the area far from the free flowing state. In Section 4, we perform clustering on temporal sequences of the network-level traffic states in order to verify capacity of the tensor factorization in unveiling the underlined large-scale traffic dynamics. Section 5 presents long-term temporal prediction results of large-scale traffic dynamics.

4. Experimental Results

As shown in Eq.3 and Eq.7, the dimensionality `r` of tensor projections needs to be fixed manually. To choose a proper value for `r`, we compute element-wise sum-of-square reconstruction error (shown in the first term of Eq.3) with different `r` from 5 to 45, as illustrated in Figure 3. The reconstruction error declines much slower when `r` is larger than 35, indicating 35 dimensional tensor projection is competent for describing underlying temporal dynamics of network-level congestion patterns. Therefore, we set `r` to be 35. To benefit from the regularization term in Eq.3 and Eq.7, we set the regularization parameters λ and λ_M to be 100. Empirically it is enough to enforce the constraint.

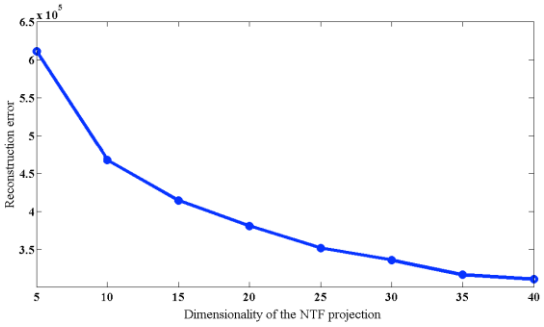


Figure 3. Reconstruction errors corresponding to different dimensionality settings for NTF projection

4.1. Clustering of temporal sequences of network-level traffic states

As described in Section 3, there are generally three settings (ITD, ATD and ETD) in simulations of traffic flow dynamics, leading to different spatial congestion patterns during peak hours. Correspondingly, we set the number of clusters in K-means to be 3. Given this configuration, we aim to find out whether the clustering results are consistent with the different demand settings we used in our set of simulations, which would illustrate the validity of the proposed tensor factorization method. Furthermore, since ATD simulation setting includes both globally light and heavy congestion level due to variances of traffic demands injected to the O-D matrix, we increase the number of clusters from 3 to 4 in order to verify whether the proposed clustering scheme can find out such detailed differences of global congestion level within the same setting of spatial congestion patterns. Besides, by increasing the number of the derived clusters, we aim to investigate the stability of the clustering results. For fair comparison, we involve clustering of spatial-temporal dynamics by performing the same K-means algorithm on the frontal slices of the tensor structured data, named as "Basic Clustering Scheme". In this scheme, each frontal slice of the tensor, corresponding to each temporal sequence of the network-level traffic state, is firstly unfolded to a (13627×48)-dimension vector in column-wise order. The resultant high-dimensional vectors are projected to a PCA subspace with the same dimensionality as the proposed tensor projection space, in order to avoid curse-of-dimensionality of K-means algorithm. The PCA projections of the frontal slices are then fed to K-means engine to get clustering memberships of the temporal sequences.

Figure 4 illustrates the cluster structures of temporal sequences illustrated in the 3D-PCA space. Figure 4(a) illustrates clustering results when the number of clusters equals to 3. The cluster labeled by blue legends consists of all 37 traffic sequences of the ITD setting. The red-labeled cluster consists of total 16 trajectories of the ETD setting. The left 2 of all 18 ETD simulations are grouped into the green-labeled cluster. 91 of total 93 trajectories in the green-labeled cluster correspond to the simulated sequences generated following the ATD setting. As we can see, the obtained three clusters of temporal sequences of network-level traffic states categorize trajectories of different traffic demand settings accurately. We name them henceforth by "Isotropic Congestion Trajectory" (ICT) and "Anisotropic Congestion Trajectory" (ACT) and "Heavy Congestion Trajectory" (HCT) respectively. By increasing the number of clusters from 3 to 4, we can find the ACT cluster is divided further into two sub-clusters, corresponding to different general congestion level during peak hour, named as "Light Anisotropic Congestion Trajectory" (LACT) and "Heavy Anisotropic Congestion Trajectory" (HACT). Figure 4(b) shows the further sub-division of the ACT cluster. The cluster LACT consists of total 41 temporal sequences, among which 39 sequences correspond to the ATD setting with low traffic demands, leading to globally low congestion level. The remaining 2 sequences correspond to the same setting with relatively high traffic demands in the O-D matrix. Among total 52 sequences in the HACT cluster, 50 sequences correspond to the setting with high traffic demands, leading to globally heavy congestion level. The minor overlapping between these two clusters (2 sequences) is caused by the smooth transition between the low and high traffic demands in the O-D matrix of

the ACT setting. As we can see in Figure 4(b), by increasing the number of clusters, our clustering scheme not only preserves the clustering structure that separates different simulation settings of traffic demands accurately, but also unveils variations of global congestion level of the same simulation setting.

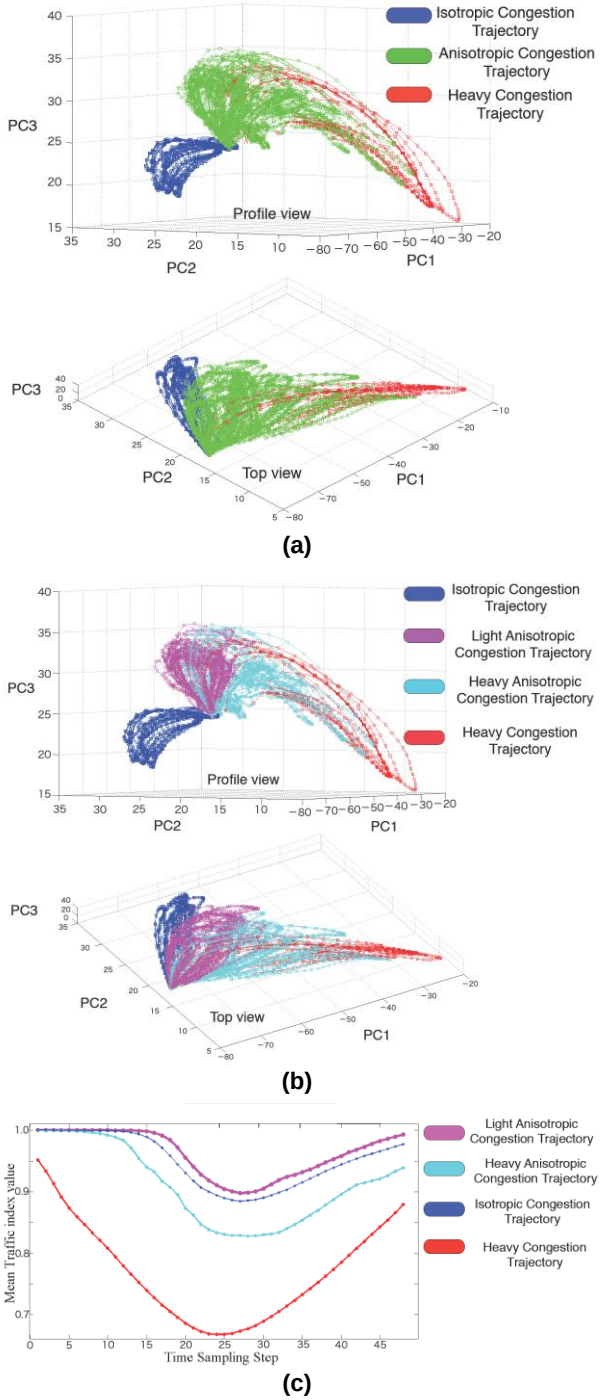


Figure 4. (a) Clustering of temporal sequences of network-level traffic states when the number of clusters is 3; (b) Clustering of temporal sequences of network-level traffic states when the number of clusters is 5; (c) Comparison of average temporal patterns of the 4 trajectory clusters.

For each time sampling step, we consider mean traffic index value (average traffic index value of all 13627 links in the network) as a crude measure of global congestion level at the each specific time period, the sequence of totally 48 mean index values in one simulation form a general evaluation of large-scale traffic dynamics of the simulation. We further calculate average of all the 48-D sequences of mean traffic index values in each cluster. The resultant 48-D average sequence represents general temporal patterns of network-level traffic states in the corresponding cluster. Figure 4(c) illustrates those average sequences of mean index values corresponding to each trajectory cluster. The cluster ICT and LACT have similar variation mode of global congestion level during the peak hour. In HCT, the network starts to suffer from congestion since the beginning of simulations, which is much different from the others and consistent with the ETD simulation setting. The cluster HACT contains an intermediate level of congestion level, compared with the cluster ICT/LACT and HCT. The proposed clustering analysis provides a divide-and-conquer solution to unveil underlined large-scale traffic dynamic patterns. Network-level traffic state sequences in the same cluster share common spatio-temporal dynamic characteristics. By analyzing the typical dynamic process of each cluster, we can improve controllability of the large-scale traffic dynamic process.

Figure 5 illustrates clustering results using Basic Clustering Scheme and compares them with those obtained using the proposed tensor factorization. The number of derived clusters is 3 for comparison. As shown in the figure, total 89 sequences in the green-labeled cluster are composed by all 39 sequences of the ITD setting, 39 sequences of the ATD setting with low traffic demands and 11 sequences of the ATD setting with similar level of traffic demands. Although they differ in the spatial settings of the traffic demands, they share similar global congestion level during the peak hours. Therefore, this cluster is noted as "Light Congestion Cluster". The blue-labeled cluster consists of all 41 sequences corresponding to the ATD setting with high traffic demands and 4 sequences generated by the ETD setting, which present intermediate level of global congestion in the network, named as "Intermediate Congestion Cluster" in the figure. The red-labeled cluster is composed by all left 14 sequences of the ETD setting containing utterly heavy congestion during the peaking hours, named as "Heavy Congestion Cluster". This clustering result only identifies variations of global congestion level over the whole network, but fails to identify spatial differences of traffic congestion between different simulation settings. Since Basic Clustering Scheme and the proposed r-NTF based method both use K-means as the final clustering toolkit, the comparison between their corresponding clustering results highlights merits of the proposed r-NTF in extracting spatio-

temporal congestion pattern. In Basic Clustering Scheme, unfolding sequences of network-level traffic states data into causes destruction of spatio-temporal correlation structures of traffic states in the network. PCA performed on the unfolded vectors only reduces their dimensions, thus cannot detect typical temporal evolution patterns of spatial congestion occurrences. In contrast, the 3-way tensor structure preserves spatio-temporal data structure and the proposed r-NTF scheme is designed to unveil spatial-temporal traffic state variation patterns directly. In this sense, the proposed r-NTF is not only a dimension reduction tool, but also a feature extraction procedure to construct a compact representation of spatial-temporal traffic state patterns.

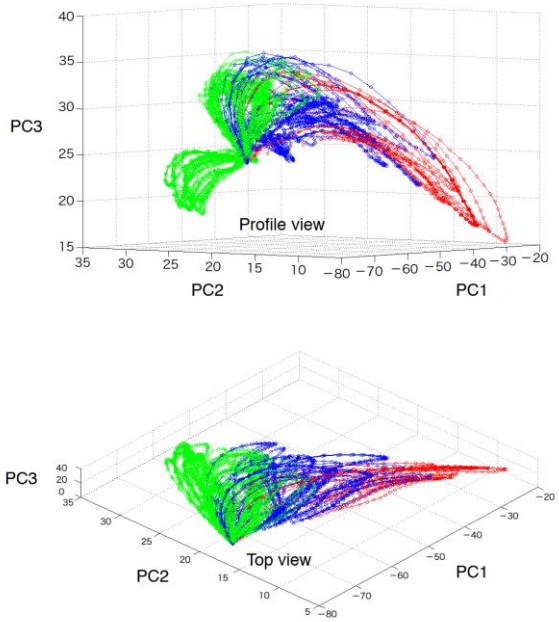


Figure 5. Clustering of temporal sequences of network-level traffic states using Basic Clustering Scheme

5. Long-term temporal prediction

In the IAURIF database, about 1/3 of the whole 13627 links in the network rarely suffer from congestion during the simulation process. They are thus out of the scope for the prediction task. To focus on the local regions that are likely to have fluctuations of traffic flow states, we select only the 3415 most congested links during the peak hours of the simulation as our prediction target. Furthermore, in temporal scale, the first 10 and last 5 sampling timesteps are the periods during which the network stays to be globally free-flowing. Therefore we turn to focus on the time period starting from the 11th steps to the 43rd steps in prediction. In these 33 sampling timesteps, we choose the first 5 as the observed part in the simulation, which covers the starting of the peak hour in each simulation. Temporal evolution during the remaining 28 timesteps are then considered as unobserved ones and used for prediction. With this setting, the prediction task can be formulated as a long-term forecast of large-scale traffic dynamics for the selected 3415 links.

We select randomly 120 of the whole 146 simulated scenes as the historic data set. The left 26 form the testing set. Both sets are rearranged into 3-way tensor structure $T^{historic} \in R^{3415*33*120}$ and $T^{test} \in R^{3415*33*26}$. Each test simulation sequence corresponds to the frontal slice of T^{test}, with the first 5 columns observed and left 28 columns for prediction. Prediction is then performed following the proposed scheme in Section 2.4. To evaluate prediction performance, we calculate the average of entry-wise Euclidean distances between the reconstructed columns in each test sequence and their ground truths. Smaller average values indicates less prediction error, thus better accuracy in estimating the large-scale dynamics. Both mean and variance of all the 26 average Euclidean distance measures are used to evaluate the general prediction performances over the whole testing set, named as "Average Prediction Error" and "Variance of Prediction Error" in this work.

Table 1. Prediction accuracies with different settings of nearest neighbors

The number of nearest neighbors	Average Prediction Error	Variance of Prediction Error
1	0.0489	0.0011
3	0.0337	0.0009
5	0.0595	0.0011
7	0.0673	0.0013
9	0.0811	0.0014

We illustrate prediction performances of different settings of nearest neighbors in Table 1. In the testing set of the IAURIF database, 3-nearest-neighbors is the best choice for predicting the unobserved traffic dynamics. By increasing the number of nearest neighbors used from 1 to 3, the average of prediction error reduces monotonically while the variance stays relatively stable. Further increase of the number of nearest neighbors leads to decline of the prediction accuracy, larger average prediction error and slightly larger variance, as shown in Table 1. On one hand, more nearest neighbors provide more heuristic information about the tendancy of unobserved temporal evolution. On the other hand, we should notice that more nearest neighbors do not ensure better prediction. The historic sequences at the end of the nearest neighboring list contains much more difference in traffic dynamics than those at the head of the list, which introduces noise into reconstruction and causes the declination.

Table 2. General Prediction Error of the three methods

The prediction schemes	Average Prediction Error	Variance of Prediction Error
Historic-Average	0.0492	0.0014
Historic-NN	0.0368	0.0011
The proposed NTF based prediction	0.0337	0.0009

To verify validity of the proposed prediction scheme, we compare the prediction performances of the proposed method with the other two baseline methods. The first one calculates average traffic index for each of 3415 links at each of the 28 time sampling steps involved in the test with respect to the corresponding time sampling step in the historic data, generating a matrix $M^{average} \in R^{3415\times 28}$ as the estimates of the unobserved traffic states. It is then used as the prediction results directly, named as "Historic-Average". In the second scheme, we follow a similar way to generate the estimates. Instead of calculating the average value with respect to all historic data, we generate it heuristically by calculating the average traffic index for each link only with respect to the nearest neighbors of the test sequence in the historic data, labeled by "Historic-NN". The number of nearest neighbors is set to be the same value as used in the tensor based prediction scheme. The comparison of prediction performances is listed in Table 2.

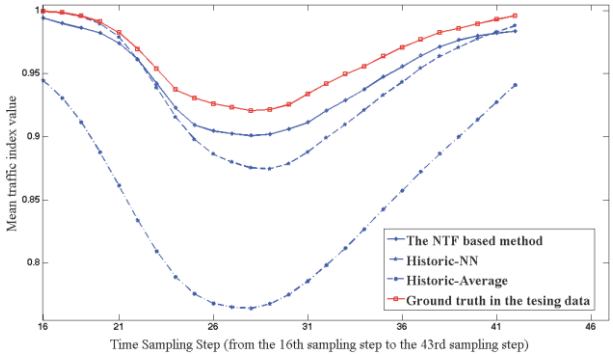


Figure 6. Comparisons of the prediction performances

Figure 6 illustrates the mean traffic index value of the total 28 time sampling steps involved in prediction in one testing time period (ranging from the 16th time sampling step to the 43rd step in the whole 48 steps). The mean traffic index values are derived from the ground-truth and the prediction results of the three methods respectively. As expected given large demand variability in the historic, Historic-NN outperforms Historic-Average to a large extent. The Historic-Average does not use any heuristic information about temporal evolution of traffic states. Purely average operation with respect to all historic data ignores the difference between traffic state variation patterns of different time periods. Compared with Historic-NN, the proposed non-negative tensor factorization based method detects the nearest neighbors more accurately in the compact tensor projection space than in the original high dimensional space. Thus, it improves further prediction performance.

6. Conclusions

In this work, we apply a tensor-factorization-based spatio-temporal data analysis scheme to investigate large-scale traffic dynamics in a urban transportation network. Our work contributes in the following aspects. Firstly, we propose to use 3-way tensor as the basic representation of temporal evolution of traffic states over the whole network. Based on the 3-way tensor structure, we use tensor factorization to generate compact signatures of spatio-temporal traffic state evolution patterns within each concerned time period. Benefited from the work in data representation level, we achieve to find out typical temporal evolutions of the global traffic states using clustering analysis. Finally, we inject K-NN based heuristic information into the tensor factorization model using a unified optimization framework. Through it, we can predict variation patterns of global traffic state configurations accurately. Although we only focus on long-term prediction of traffic dynamics, the proposed tensor reconstruction can also be easily extended to estimate missing observations of traffic states due to faults of sensors and noise in the GPS signals.

7. Acknowledgments

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

References

[1] Herring R., Hofleitner A., Amin S., Nasr T., Khalek A., Abbeel P., and Bayen A.: 'Using mobile phones to forecast arterial traffic through statistical learning', Proc. 89th Transportation Research Board Annual Meeting, Washington D.C. (USA), January 2010.

[2] Kinuta Y., Kitamura S., Nakamura T., Makimura K., Takahashi M. and Morokawa T.: 'Examination of the Applicability of Probe Car Data to Assessment of the Effects of Link-Safety Projects', International Journal of Intelligent Transportation Systems Research, vol.8, issue 2, pp.67-76, 2010.

[3] Work D., Blandin S., Tossavainen O., Piccoli B., and Bayen A.: 'A traffic model for velocity data assimilation', Applied Mathematics Research eXpress (AMRX), 1, pp.1-35, 2010.

[4] Thiagarajan A., Sivalingam L., LaCurts K., Toledo S., Eriksson J., Madden S., and Balakrishnan H.: 'VTrack: Accurate, Energy-Aware Traffic Delay Estimation Using Mobile Phones', Proc. 7th ACM Conf. Embedded Networked Sensor Systems (SenSys), Berkeley CA (USA), pp.85-98, November 2009.

[5] Krause A., Horvitz E., Kansal A., and Zhao F.: 'Toward community sensing', Proc. of Int. Conf. on Information Processing in Sensor Networks (IPSN), St Louis (USA), pp.481-492, April 2008.

[6] Liu H., and Ma W.: 'A virtual vehicle probe model for time-dependent travel time estimation on signalized arterials', Transportation Research Part C, 17 (1), pp.11-26, 2009.

[7] Jaeger A., Bibmeyer N., Stubing H. and Huss A.S.: 'A Novel Framework for Efficient Mobility Data Verification in Vehicular Ad-hoc Networks', International Journal of Intelligent Transportation System Research, vol.10, issue 1, pp.11-21, 2012.

[8] Vanajakshi L., Subramanian S.C., and Sivanandan R.: 'Travel time prediction under heterogeneous traffic conditions using global positioning system data from buses', IET Intelligence Transportation System, 3 (1), pp.1-9, 2009.

[9] Chowdhury D., Santen L., and Schadschneider A.: 'Statistical physics of vehicular traffic and some related systems', Physics Report, 329, pp.199-329, 2000.

[10] Herty M., Klar, A., and Pareschi, L.: 'General kinetic models for vehicular traffic flow and Monte Carlo methods', Computational methods in applied mathematics, 5 (2), pp.155-169, 2005.

[11] Rakha H.: 'Validation of Van Aerde's Simplified Steady-state Car-following and Traffic Stream Model', Transportation Letters: The International Journal of Transportation Research, 1 (3), pp. 227-244, 2009.

[12] Nagel K., and Schreckenberg M.: 'A cellular automaton model for freeway traffic', Journal of Physics, 2, pp. 2221–2229, 1992.

[13] Blandin S., Work D., Goatin P., Piccoli B., and Bayen A.: 'A general phase transition model for vehicular traffic', SIAM Journal on Applied Mathematics, to appear, 2011.

[14] Wang Y., and Papageorgiou M.: 'Real-time freeway traffic state estimation based on extended kalman filter: a general approach', Transportation Research part B, 39, pp.141–167, 2005.

[15] Statthopoulos A., and Karlaftis M.G.: 'A multivariate state space approach for urban traffic flow modeling and predicting', Transportation Research Part C, 11, pp.121–135, 2003.

[16] Helbing D., Rarita L., D'Apice C., and Piccoli B.: 'Sensitivity analysis of permeability parameters for flows on Barcelona networks', Journal of Differential Equations, vol.249, n°12, pp.3110-3131, 2010.

[17] Helbing D.: 'Derivation of non-local macroscopic traffic equations and consistent traffic pressures from microscopic car-following models', European Physical Journal B, vol.69, n° 4, pp.539-548, 2009.

[18] Helbing D., and Johansson A.: 'On the controversy around Daganzo's requiem for and Aw-Rascie's resurrection of second-order traffic flow models', European Physical Journal B, vol.69, n° 4, 549-562, 2009.

[19] Ghosh B., Basu B., and O'Mahony M.: 'Multivariate short-term traffic flow forecasting using time-series analysis', IEEE Transaction on Intelligence Transportation Systems, 10 (2), pp.246 – 254, 2009.

[20] Yin H., Wong S.C., Xu J., and Wong C.K.: 'Urban traffic flow prediction using a fuzzy-neural approach', Transportation Research Part C: Emerging Technologies, 10 (2), pp.85–98, 2002.

[21] Quek Y., Pasquier M., and Lim B.: 'POP-TRAFFIC: A Novel Fuzzy Neural Approach to Link Traffic Analysis and Prediction', IEEE Transactions on Intelligent Transportation Systems, 7 (2), pp.133–146, 2006.

[22] Vlahogianni E. I.: 'Enhancing Predictions in Signalized Arterials with Information on Short-Term Traffic Flow Dynamics', Journal of Intelligent Transportation Systems, 13(2), pp. 73 – 84, 2009.

[23] Furtlehner C., Lasgouttes J.-M., and de La Fortelle A.: 'A belief propagation approach to traffic prediction using probe vehicles', Proc. 10th Int. Conf. Transportation Systems (ITSC), pp.1022–1027, 2007.

[24] Arel I., Liu C., Urbanik T., and Kohis A.G.: 'Reinforcement learning-based multi-agent system for network traffic signal control', IET Intelligence Transportation System, 4 (2), pp.128-135, 2010.

[25] Li S. Z.: 'Markov Random Field Modeling in Image Analysis' (Springer Press, 2009, 3rd edn)

[26] Shoham Y. and Leyton-Brown K.: 'Multiagent Systems: Algorithmic, Game- Theoretic, and Logical Foundations' (Cambridge University Press, 2008)

[27] Geroliminis N., and Daganzo C.F.: 'Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings', Transportation Research Part B: Methodological, vol.42, num.9, pp.759-770, 2008.

[28] Geroliminis N. and Sun J.: 'Properties of a well-defined Macroscopic Fundamental Diagram for urban traffic', Transportation Research Part B Methodological, vol.45,no.3,pp.605-617,2011.

[29] Ji Y. and Geroliminis N.: 'On the spatial partitioning of urban transportation networks', Transportation Research Part B Methodological, vol. 46, N° 10, p. 1639-1656, 2012.

[30] Han Y., and Moutarde F., 'Statistical Traffic State Analysis in Large-scale Transportation Networks Using Locality-Preserving Non-negative Matrix Factorization',

IET Journal of Intelligent Transportation System, vol.7, issue 3, pp.283-295, 2013.

[31] Kanungo T., Mount D.M., Netanyahu N.S., Piatko C.D., and Wu A.Y., 'An efficient k-means clustering algorithm: Analysis and implementation', IEEE Trans. on Pattern Analysis and Machine Intelligence, 24, pp.881–892, 2002.

[32] Chi C.E. and Kolda T.G.: 'On Tensors, Sparsity, and Nonnegative Factorizations', SIAM Journal on Matrix and Applications, vol.33, n° 4, pp.1272-1299, 2012.

[33] Kolda T.G. and Bader B.W.: 'Tensor Decomposition and Applications', SIAM Review, vol.51, n° 3, pp.455-500, 2009.

[34] Liu J., Liu J., Wonka P., and Ye J.P.: 'Sparse Non-negative Tensor Factorization Using Columnwise Coordinate Decent', Pattern Recognition, vol.45, n° 1, 2012.

[35] Spiegel S., Clausen J., Albayrak S, and Kunegis J.: 'Link prediction on evolving data using tensor factorization', proc. 15th international conference on New Frontiers in Applied Data Mining, pp.100-110, 2011.

[36] Lee D.D. and Seung H.S.: 'Algorithms for non-negative matrix factorization', Proc. 13th Neural Information Processing Systems (NIPS), Denver (USA), pp.556-562, 2000.

[37] Xu, W., Liu, X., and Gong, Y.H.: 'Document clustering based on non-negative matrix factorization', Proc. 26th ACM SIGIR, Toronto (Canada), pp.267-273, 2003.

[38] Wang Y., Jiang Y., Wu Y., and Zhou Z.H.: 'Local and Structural Consistency for Multi-Manifold Clustering', Proc. 22nd International Joint Conference on Artificial Intelligence (IJCAI), Barcelona (Spain), pp.1559-1564, 2002.

[39] Chen G.L., and Lerman G.: 'Spectral curvature clustering', International Journal of Computer Vision, 81 (3), pp.317–330, 2009.

[40] Chung F.R.K.: 'Spectral Graph Theory', CBMS Regional Conference Series in Mathematics, 1997.

[41] Agrawal R., Gehrke J., Gunopulos D.: 'Automatic Subspace Clustering of High Dimensional Data', Data Mining and Knowledge Discovery, vol.11, issue 1, pp.5-33, 2005.

[42] Boyd S. and Vandenberghe L.: 'Convex Optimization', Cambridge University Press.

[43] Marchal F.: 'Contribution to dynamic transportation models', PhD Thesis, University of Cergy-Pontoise, 2001.

[44] De Palma A. and Marchal F.: 'Real cases applications of the fully dynamic METROPOLIS tool-box: an advocacy for large-scale macroscopic transportation systems', Networks and Spatial Economics, 2 (4), pp. 347–369, 2002.

[45] Jolliffe I.T.: 'Principal Component Analysis' (Springer Press, 2002, 2nd edn)