IET Intelligent Transport Systems

Research Article

Traffic flow data compression considering burst components

ISSN 1751-956X
Received on 28th November 2016
Revised 18th July 2017
Accepted on 11th August 2017
E-First on 28th September 2017
doi: 10.1049/iet-its.2016.0328
www.ietdl.org

Shuo Feng[1], Ruimin Ke[2], Xingmin Wang[3], Yi Zhang[1], Li Li[1] ✉
[1]*Department of Automation, Tsinghua University, Beijing 100084, People's Republic of China*
[2]*Department of Civil and Environmental Engineering, University of Washington, Seattle, WA 98195, USA*
[3]*Department of Automotive Engineering, Tsinghua University, Beijing 100084, People's Republic of China*
✉ *E-mail: li-li@tsinghua.edu.cn*

Abstract: Many recent applications of intelligent transportation systems require both real-time and network-wide traffic flow data as input. However, as the detection time and network size increase, the data volume may become very large in terms of both dimension and scale. To address this concern, various traffic flow data compression methods have been proposed, which archive the low-dimensional subspace rather than the original data. Many studies have shown the traffic flow data consist of different components, i.e. low-dimensional intra-day trend, Gaussian type fluctuation and burst components. Existing compression methods cannot compress the burst components well and provide very limited choices of compression ratio (CR). A better compression method should have the ability to archive all the dominant information in different components of traffic flow data. In this study, the authors compare the influence of different data reformatting, archive the bursts defined before in descending order with respect to the absolute value of the burst points and propose a flexible compression framework to balance between burst components and low-dimensional intra-day trend. Experimental results show that the proposed framework promotes the reconstruction accuracy significantly. Moreover, the proposed framework provides more flexible choices with respect to CR, which can benefit a variety of applications.

1 Introduction

Intelligent transportation systems have been playing a great role in different practical and research tasks such as traffic control strategy, remote sensing, route guidance and traffic prediction [1–6]. Recently, more sophisticated applications in these fields have become available with the emergence and availability of the large amount of traffic flow data, many of which are in real time and network wide [2, 4]. Large roadway networks are typically composed of thousands of road segments, and traffic flow data of each segment are normally recorded as one data dimension in the dataset. For data acquisition systems that have temporal resolution of 30 s [5], approximately one million data points are generated by a single sensor annually. That is to say, traffic flow data is becoming both high dimensional and large scale, which may easily cause the curse of dimensionality for many data-driven applications and put strain on resources of data management systems [7, 8]. Traffic flow data compression, as one of the most effective solutions, has been proposed to address these kinds of problems [9].

Many studies have shown that there is an inherent low-dimensional subspace underlying the high-dimensional raw data, which is able to interpret most information of the original traffic flow data matrix [7, 9]. The fundamental idea of data compression is to archive the low-dimensional subspace by the similarity of the data rather than the original data, based on which the reconstruction can be conducted. There have been many methods to archive the low-dimensional subspace of the traffic flow data such as traffic data compression methods based on artificial neural network [10], wavelet transform [11], principal component analysis (PCA) [7], Kronecker product (KP) [12] and tensor [13, 14]. Note that the principal theory of traffic flow data compression methods is different from that of traditional compression methods such as those based on Morse coding [15], Shannon theorem [16] and Huffman coding [17]. As Fig. 1 shows, a data compression procedure can be divided into two stages, the first of which is based on the redundancy in the raw data and get the low-dimensional data, the second of which is based on the redundancy in the coding of the files [18, 19] and get the computer code. In this paper, we focus on the first stage compression based on the instinct structure of traffic flow time series data.

However, as shown in [20–22], the traffic flow data are heterogeneous and consist of different components, namely the low-dimensional intra-day trend, Gaussian type fluctuation and burst components. The existing compression methods cannot compress the burst components well and the reconstruction error in such data points is significantly large. Furthermore, the methods can only provide limited choices of compression ratio (CR) which can hardly meet the practical requirements. Therefore, it is reasonable to come up with the hypothesis that the superior compression method should archive all the dominant information of all the three components of traffic flow data, rather than just the low-dimensional subspace. In this paper, we proposed a flexible compression framework to implement the compression with burst components considered. To make the research concise and focus on the key contributions, we mainly discuss the burst components compression in this paper. More detailed information about fluctuation analysis can be found in [23, 24].

Experimental results show the proposed compression framework promotes the reconstruction accuracy significantly and provides more flexible choices of CR. Further studies show that though burst components or outliers are the interferences to the data analysis in many research areas, burst components of traffic flow data contain crucial information of non-recurrent traffic breakdowns caused by traffic accidents, adverse weather or big events.

To implement the proposed framework, three essential problems need to be addressed. The first problem is how to reformat the raw data, so that they are able to be used by the algorithm; also, properly reformatting the data is important to the fairness of comparison of different algorithms. The second problem is how to archive the principal information of burst components, and the third one is how to balance between burst components and low-dimensional intra-day trend. Fig. 2 shows the four steps to implement the framework and their corresponding effects. In the second step of the framework, the matrix decomposition methods

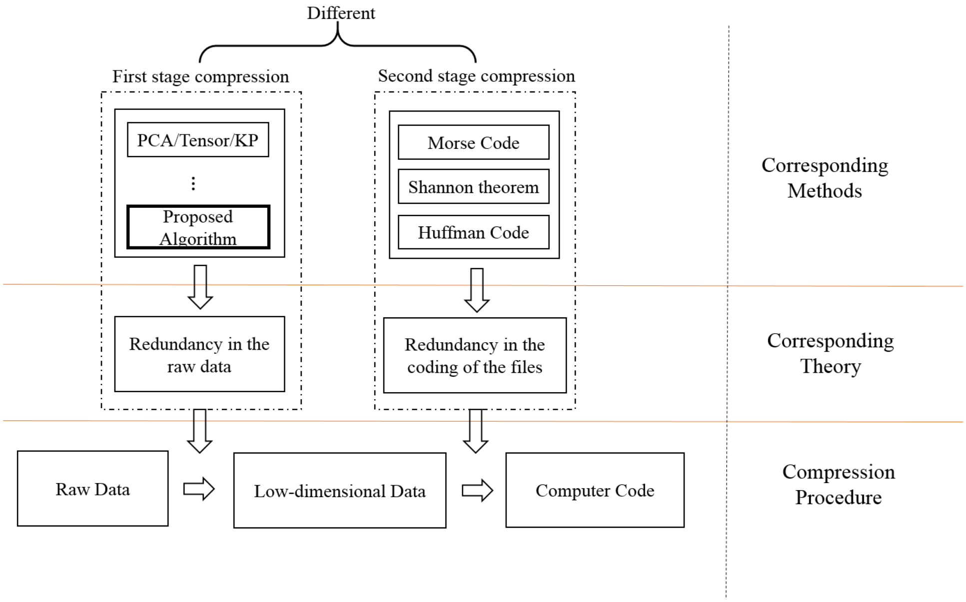

Fig. 1 *Comparison between two different compression stages*

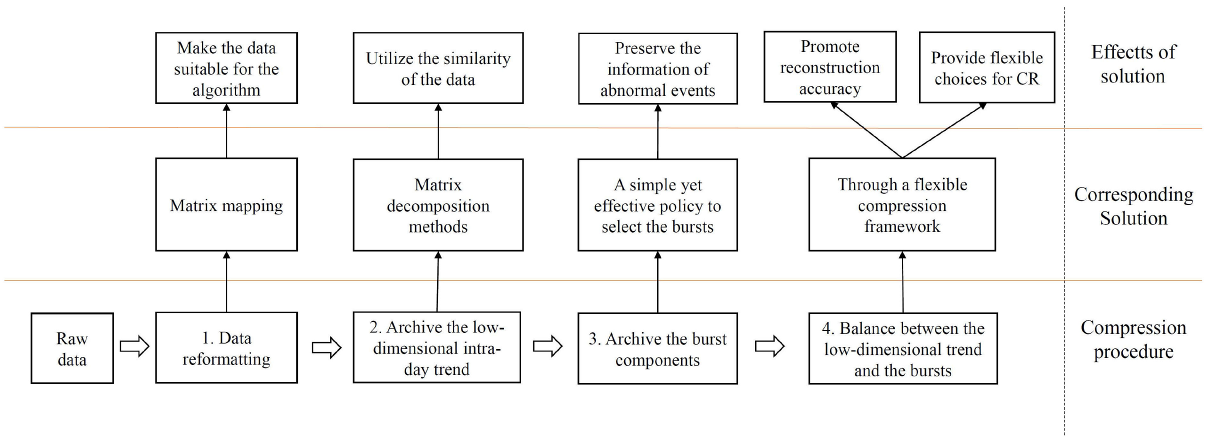

Fig. 2 *Flowchart of the proposed compression framework*

to archive the low-dimensional intra-day trend have been well-developed in previous studies. In our paper, three most widely used matrix decomposition methods which are based on PCA [7], KP [12] and tensor [13, 14] are adopted. The overall processing pipeline is actually developed based on addressing the aforementioned three problems in an effective manner. For the first problem, we compare the influence of different ways of matrix mapping for different amounts of traffic networks. For the second problem, we employ a simple yet effective policy: archiving the bursts defined before in descending order with respect to the absolute value of the burst points. The sparse matrix of the burst points can be reconstructed by a series of 2-tuple, which are composed by a numerical value term and a location term obtained by lexicographical recording of the sparse matrix. For the last one, we propose a flexible compression framework to balance between burst components and low-dimensional intra-day trend.

The rest of this paper is organised as follows. Section 2 explains the data sources and performance indexes. Section 3 explains why and how to implement the proposed flexible compression framework. The experimental test and discussion are presented in Section 4. Finally, the conclusions are summarised in Section 5.

2 Data sources and performance indexes

In this paper, we select the flow data of traffic networks from the publicly accessible performance measurement system dataset [5]. The sampling interval in this work is 5 min. The traffic networks are all within the state of CA in the USA, and the traffic flow data have a time range from 1st August 2011 to 21st August 2011. To compare the influence of the sizes of traffic networks, we consider three test networks at small, middle and large size. The small test network has ten traffic flow detectors (Category A (CATA)), the middle one has 100 detectors (Category B (CATB)) and the large one has 1000 detectors (Category C (CATC)). The data volumes of the test datasets are 60,480 (CATA), 604,800 (CATB) and 6,048,000 (CATC), respectively, in the unit of bytes.

Suppose that we have m detectors and each detector records n consecutive points of traffic flow data, we could get the raw data matrix $\boldsymbol{X} \in R^{m \times n}$ as

$$\boldsymbol{X} = [X(1), X(2), \ldots, X(m)]^{\mathrm{T}}, \tag{1}$$

where $X(i) = [x_i(1), X_i(2), \ldots, X_i(n)]^{\mathrm{T}}$ represents the data collected by the i detector. According to our paper, it is simple enough to present the data in matrix. For example, when we want the value of the i point in j day of k detector, $v(i, j, k)$, we have

$$v(i, j, k) = \boldsymbol{X}(k, 288 * (j-1) + i), \tag{2}$$

where 288 is the total number of points in 1 day. Therefore, we can just present the data in matrix.

The most crucial objectives for a compression method is to (i) reach as high reconstruction accuracy as possible and (ii) use as little storage space as possible. Therefore, two indexes are normally adopted to evaluate the performance of a compression method.

One is the CR to measure the amount of storage space as

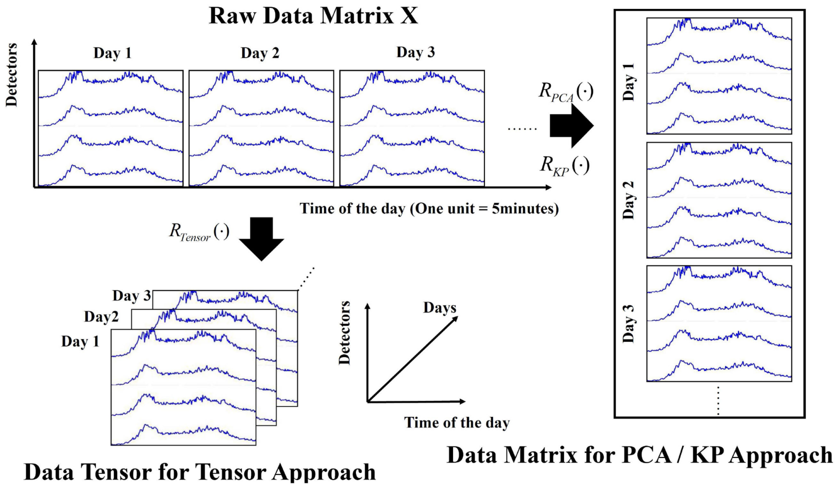

Fig. 3 *Illustration of the data reformatting process*

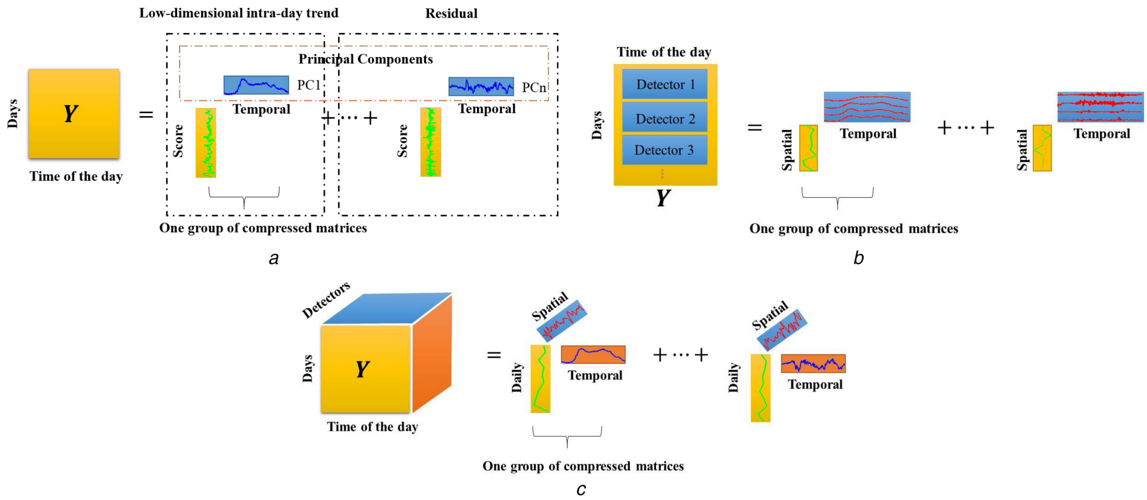

Fig. 4 *Illustration to help clarify the relationship between the matrix decomposition methods and data compression*
(a) PCA-based, ***(b)*** KP-based, ***(c)*** Tensor-based

$$\mathrm{CR} = \frac{\text{bytes of compressed data}}{\text{bytes of raw traffic data}}. \tag{3}$$

Another index is the relative absolute deviation (RAD) to measure the reconstruction accuracy as

$$\mathrm{RAD}(\boldsymbol{X}, \tilde{\boldsymbol{X}}) = \frac{1}{\| X \|} \sum_{i=1}^{m} \sum_{j=1}^{n} |\tilde{x}_i(j) - \tilde{x}_i(j)|, \tag{4}$$

where $\| X \| = \sum_{i=1}^{m} \sum_{j=1}^{n} |\tilde{x}_i(j)|$ and $\tilde{X} \in R^{m \times n}$.

3 Flexible compression framework

3.1 Data reformatting

To make the raw data suitable for the algorithm, a data reformatting operator $R(\cdot)$ is considered to map $X \in R^{m \times n}$ into matrices or tensors. To preserve the instinct similarity of traffic flow data, we should choose a data reformatting operator that has practical meaning such as hours, days and weeks. Taking Fig. 3 as an example, we format the data collected by one detector in 1 day as a row of data matrix and data collected by one detector as a slide for data tensor.

3.2 Matrix decomposition methods

In this paper, we consider three matrix decomposition methods to archive the low-dimensional intra-day trend based on PCA [7], KP [12] and tensor [13, 14]. To help clarify the relationship between the matrix decomposition methods and data compression, we introduce the methods as the constrained optimisation problems, given by (4)–(6)

$$\min_{Y} \| R_{\mathrm{PCA}}(\boldsymbol{X}) - Y \| \quad \text{s.t. } Y = U \cdot V = \sum_{i=1}^{r} u_i \cdot v_i^{\mathrm{T}}, \tag{5}$$

$$\min_{Y} \| R_{\mathrm{KP}}(\boldsymbol{X}) - Y \| \quad \text{s.t. } Y = \sum_{i=1}^{r} U_i \otimes V_i, \tag{6}$$

$$\min_{Y} \| R_{\mathrm{tensor}}(\boldsymbol{X}) - Y \| \quad \text{s.t. } Y = \sum_{i=1}^{r} \lambda_i u_i^1 \circ u_i^2, \ldots, \circ u_i^N, \tag{7}$$

where $\| \cdot \|$ is a norm defined specifically by different methods and Y is the aimed matrix (tensor). Here, $\cdot$ denotes for matrix product, $\otimes$ denotes KP and $\circ$ denotes tensor product. $\boldsymbol{u}_i$ is the ith column vector of U and so is v_i. λ_i is a scalar. $u_i^s = (u_{i,1}^s, u_{i,2}^s, \ldots, u_{i,I}^s)^{\mathrm{T}}$. I are the orders of the tensor. Note that we use the canonical polyadic decomposition here [13, 14].

Fig. 4 shows an illustration of matrix decomposition methods. As shown in (4)–(6), the matrix decomposition methods resolve the raw data matrix into summation of the matrix product of the compressed matrices. To store the appropriate groups of compressed matrices to archive the low-dimensional intra-day trend rather than the original data, we could reach the goal of compression. The remaining groups of compressed matrices are defined as the residual (see Fig. 4*a*).

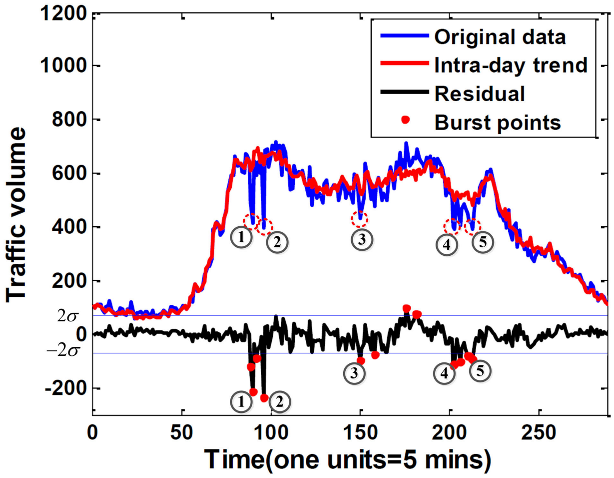

Fig. 5 *Illustration of the heterogeneous components of traffic flow data by matrix decomposition methods. We take the method based on PCA, for example, which results are similar to methods based on tensor and KP. The original data is gathered by detector (ID 311974) on 2 October 2015. The intra-day trend is calculated based on traffic flow data of five consecutive since 2 October 2015*

Table 1 Algorithm of the flexible compression framework

in: raw data matrix $\boldsymbol{X}$, the required CR

out: a combination of the compressed matrices and burst points to store

1. *data rearrangement:* $\tilde{X} = R(X)$

2. *matrix decomposition:* Resolve $\tilde{X}$ into summation of matrix product of the compressed matrices, shown as in (4)–(6)

3. *set* $i = 1$, $k = 1$

4. *Repeat*

• store the *i*th group of compressed matrices and update the residual (Fig. 4)

• get the current burst components from the residual

• sort the current burst points following the sequence that descends the absolute value of burst points:

• *set* $j = 1$.

• *repeat*

store the *j*th burst points as a 2-tuple composed by a numerical value term and a location term obtained by lexicographical recording of the sparse matrix

$$\mathrm{CCR}(k) = \frac{\text{bytes of a group of compressed matrices} \times i + j \times 2}{\text{bytes of raw traffic data}}$$

set $j = j + 1$, $k = k + 1$

• ***until*** all the current burst components are stored

• *compute* $\mathrm{CR}(i) = \dfrac{\text{bytes of a group of compressed matrices} \times i}{\text{bytes of raw traffic data}}$

• *set* $i = i + 1$

5. *until* $\mathrm{CR}(i) > \mathrm{CR}$ or all the compressed matrices are stored

6. *search the best combination***:** search the best reconstruction accuracy combination whose CCR is equal to CR within a tolerance range

3.3 Burst components compression

Fig. 5 is an illustration of the three heterogeneous components of traffic flow data. The red line denotes the intra-day trend which is archived by the matrix decomposition method based on PCA. The black line is the residual which cannot be reconstructed by the intra-day trend. The definition of residual can also see Fig. 4*a*. Suppose the standard deviation is obtained from the square root of the variance for the entire residual time series, and then we can define a point as burst point if its deviation is larger than twice the standard deviation as defined in [20–22], which are denoted by the red points. The sensitivity analysis of the threshold will be given in Section 4.4. Five burst points with large variances in Fig. 5 are selected. It is found that these burst components cannot be retrieved by matrix decomposition methods and the reconstruction errors in such points are very large.

The problem here is how to archive the principal information of the burst components. In this paper, we employ a simple yet effective policy: archiving the bursts defined before in descending order with respect to the absolute value of the burst points. The sparse matrix of the bursts can be reconstructed by a series of 2-tuples, which are composed by a numerical value term and a location term obtained by lexicographical recording of the sparse matrix.

3.4 Flexible compression framework

We have already archived the low-dimensional subspace of traffic flow data by matrix decomposition methods as well as the burst components by a simple yet effective policy. The remaining problem is how to balance between the burst components and the low-dimensional intra-day trend. To this end, we propose a flexible compression framework. The fundamental idea of the framework is to compare the storage cost and reconstruction accuracy decrease and select a better combination of the burst components and the low-dimensional intra-day trend. The algorithm details of the framework are presented in Table 1.

4 Numerical test and discussion

4.1 Influence of data reformatting

Data reformatting enables the raw data to be used by the algorithm. Different methods of reformatting will affect the performance of

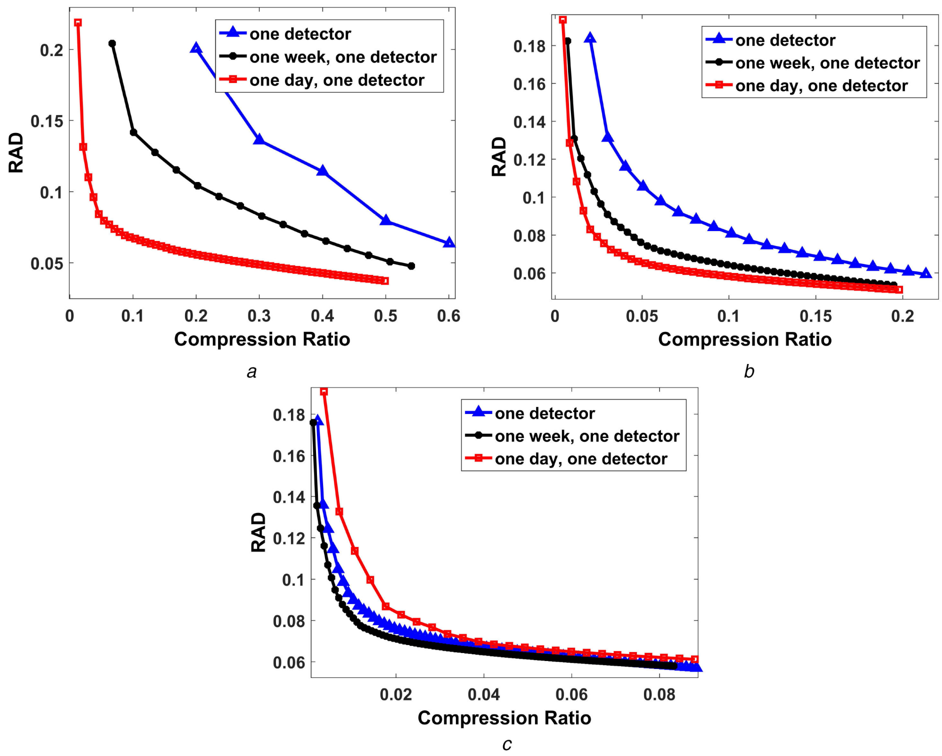

Fig. 6 *Influence of raw data matrix rearrangement for compression method based on PCA. The test network is* ***(a)*** CATA, ***(b)*** CATB, ***(c)*** CATC

compression methods. For the fairness of the succeeding experiments, the influence of data reformatting should be studied.

In this section, we discuss the influence of data reformatting on compression methods and take the method based on PCA, for example, which is similar to methods based on tensor and KP. We select data collected by one detector through all the periods (shown as 'one detector'), one detector through 1 week (shown as '1 week, one detector') and one detector through 1 day (shown as '1 day, one detector') as a row of data matrix. The results are plotted in Fig. 6. The test network is (a) CATA, (b) CATB and (c) CATC.

It can be observed that there is no constant optimal selection of data reformatting for all different sizes of traffic networks. In essence, each type of data reformatting preserves some type of similarity but yields discrepancy simultaneously and it is the balance that really matters. Take the reformatting shown by the red line in Fig. 6, for example. The matrix decomposition methods could utilise the similarity but yield the discrepancy among different detectors and days. When the network is small where the similarity plays a dominant role such as in Figs. 6*a* and *b*, the data reformatting contributes to archive the low-dimensional subspace of traffic flow data. However, when the network is large enough where the discrepancy becomes the dominant interference such as in Fig. 6*c*, the data reformatting does not perform well. For the fairness of comparison, we will take the local optimum format for different algorithms at different networks in the experiments.

4.2 Comparison of matrix decomposition methods

In this section, we compare the performance of different matrix decomposition methods based on PCA, tensor and KP at different sizes of networks. As shown in Fig. 7, method based on tensor performs better at small CR, especially when the CR is no larger than 0.02, which is consistent with the finding in our previous paper [9].

There is an instructive phenomenon that the difference among matrix decomposition methods is not obvious at large CR, especially when the size of network gets larger, as shown in Fig. 7*c*. More specifically, as the CR becomes larger, the reconstruction accuracy is not as sensitive to the matrix decomposition method, especially when the size of network gets larger. This actually implies that the main low-dimensional intra-day trend has been archived by the matrix decomposition methods, so the major errors of data compression are not in low-dimensional subspace. According to our hypothesis, the major errors come from the burst components.

4.3 Results of the flexible compression framework

In this section, we analyse the results of the proposed compression framework. First, we focus on the procedure of the framework and prove that the new framework can provide more flexible choices for CR. Then, we compare the compression performance of the proposed method with the matrix decomposition methods to confirm the hypothesis we proposed in this paper. Finally, we discuss the phenomenon of reconstruction error decrease.

Fig. 8 is an illustration of the flexible compression framework when the required CR is 0.02, where the blue triangles express the performance of the matrix decomposition method based on PCA and the red points show the possible combinations of low-dimensional intra-day trend and burst components. As shown in this figure, the flexible framework takes three steps to get the compression solution with the required CR, while the decomposition method based on PCA can only provide compression options whose CR is 0.0165 or 0.0247. Therefore, the proposed flexible compression framework provides more flexible choices for CR.

Fig. 9 shows the compression performance of the proposed method and the matrix decomposition based on PCA. The red points show the possible combinations of low-dimensional intra-day trend and burst components. The fundamental idea of the flexible compression framework is to choose the best combinations, shown as the black line. Results show that the proposed compression framework promotes the reconstruction accuracy significantly for all three test networks, especially when the CR is larger than 0.1, which confirm the hypothesis we proposed in this paper: the superior compression method should archive all the dominant information in the different components of traffic flow data, not only the low-dimensional subspace.

Fig. 10 is the reconstruction error decrease by the proposed framework at CATC. We divide this figure into three zones. In zone-I, the low-dimensional intra-day trend plays dominant role, so the reconstruction error decrease caused by the proposed method is nearly zero. In zone-II, it can be seen that the proposed framework promotes the compression performance observably, the maximum

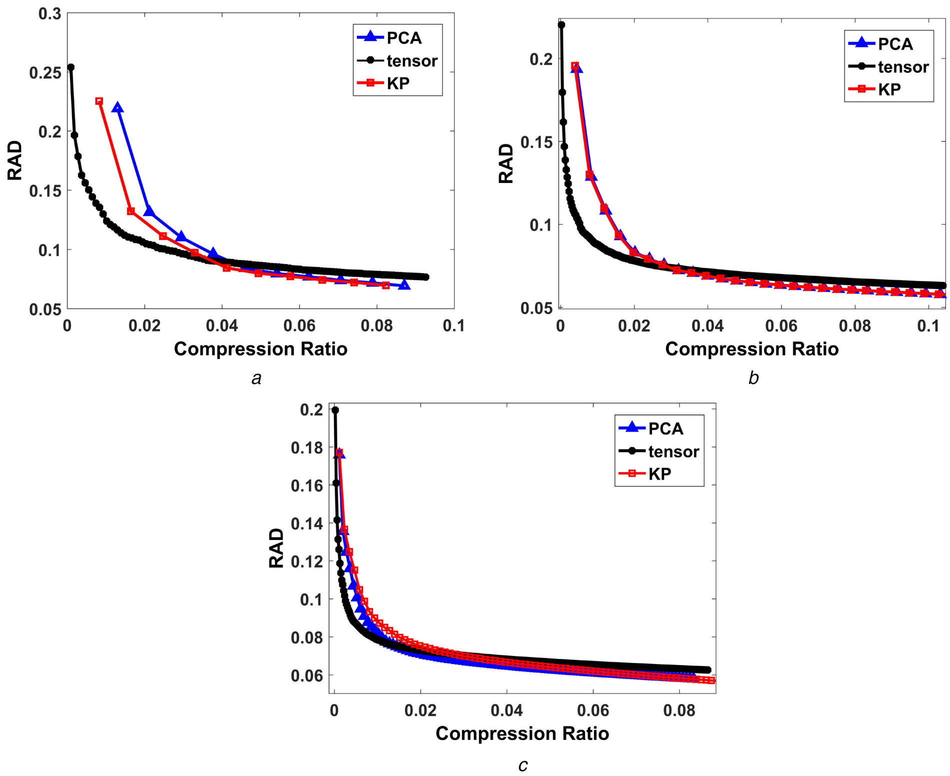

Fig. 7 *Comparison of matrix decomposition methods based on PCA, tensor and KP. The test network is* ***(a)*** CATA, ***(b)*** CATB, ***(c)*** CATC

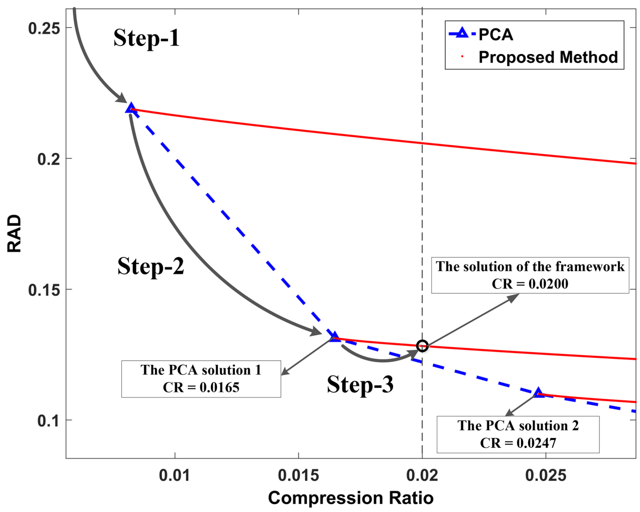

Fig. 8 *Illustration of the flexible compression framework when the required CR is 0.02. As shown in this figure, the flexible framework takes three steps to get the compression solution with the required CR, while the decomposition method based on PCA can only provide compression options whose CR is 0.0165 or 0.0247. Therefore, the proposed flexible compression framework provides more flexible choices for CR*

of which reaches to 23.2%. According to our hypothesis, the burst components play the dominant role in this zone, which means the proposed method performs well. In zone-III, the reconstruction error decrease changes very slowly and maintains at a relatively certain level. It is supposed that the low-dimensional intra-day trend and the burst components have been archived by the proposed compression method and the major errors remain in the Gaussian type fluctuation. Therefore, future efforts can be put to study the fluctuation characteristics and archive the dominant information of the fluctuation to further promote the compression performance.

4.4 Sensitivity analysis and the procedure for real situation

In our framework, the main parameter is the threshold to define a burst point, sensitivity analysis of which is given in this section. As shown in Fig. 11, we vary the threshold from 0 to 3 σ. Results show that though the proposed framework with all different threshold works better than PCA method (threshold is 0), the threshold affects the performance of the proposed framework especially when the network is middle or large (see Fig. 11*b* and *c*). Moreover, the threshold with σ has the best performance for all small, middle and large networks. Therefore, we can take σ as the threshold to define a burst point.

Inspired by the intelligent thresholds separator in [25, 26] and the results of sensitivity analysis, we can define a flexible method to separate burst points from other points by taking σ as the threshold to define a burst point in real situation. The procedure of the proposed compression framework in real situation is explained below: first, we confirm the CR as the requirement in real situation. Second, we reformat the raw data according to the results of Section 4.1 and make matrix decomposition. Third, we calculate the RAD-CR figure as Fig. 9, in which we take σ as the threshold to define a burst point and store the burst points by a series of 2-

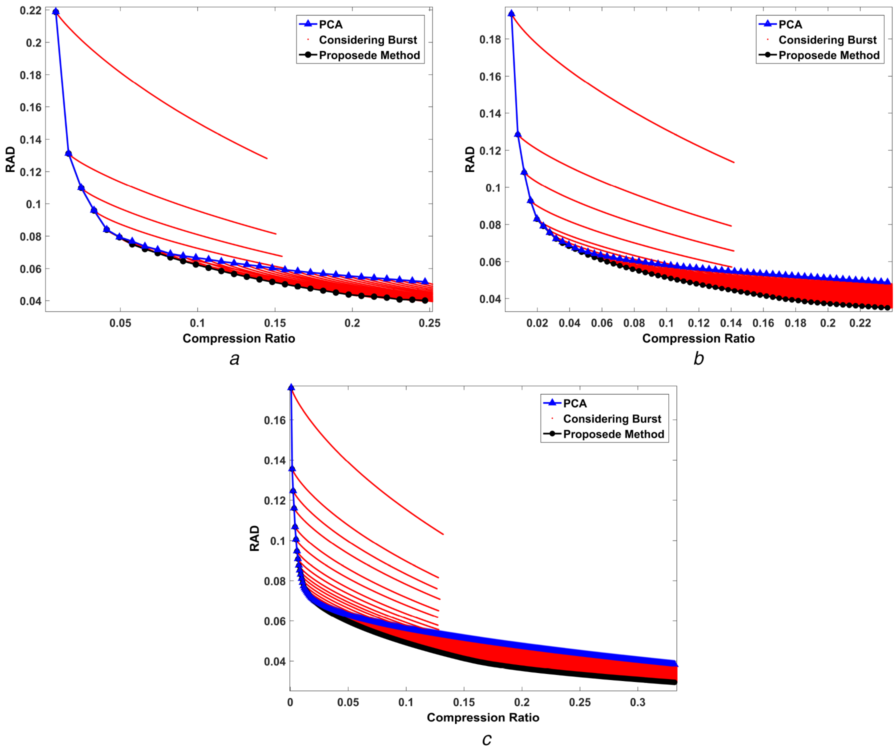

Fig. 9 *Compression performance of the proposed compression framework and the matrix decomposition based on PCA. The test network is* ***(a)*** CATA, ***(b)*** CATB, ***(c)*** CATC. Results show that the proposed compression framework promotes the reconstruction accuracy apparently for all three test networks, especially when the CR is larger than 0.1, which confirm the hypothesis we proposed in this paper

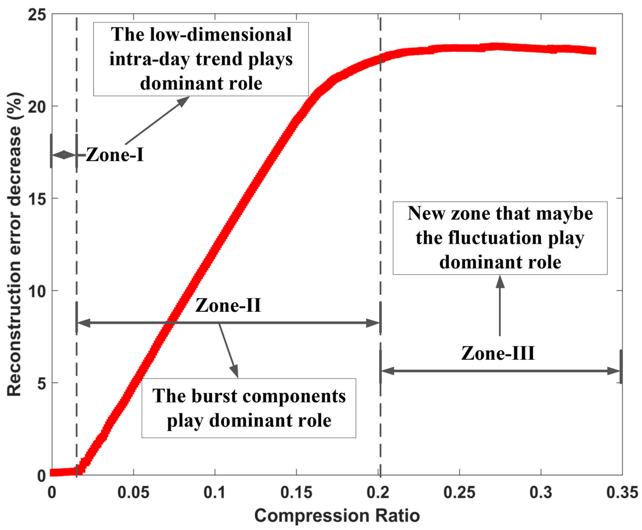

Fig. 10 *Reconstruction error decrease by the proposed method at CATC. The proposed framework promotes the compression performance observably, the maximum of which reaches to 23.2%. We can divide this figure into three zones as shown in this figure, where the low-dimensional intra-day trend plays dominant role in zone-I, the burst components play dominant role in zone-II, and zone-III is a new zone where may be the fluctuation play dominant role according to our hypothesis*

tuples which are composed by a numerical value term and a location term obtained by lexicographical recording of the sparse matrix. Finally, we get the optimal combination to implement the compression procedure as shown in Table 1.

5 Conclusion

In this paper, we propose a hypothesis that the superior compression method should archive all the dominant information in the three different components of traffic flow data. We focus on the compression considering the burst components in this work. Results confirm the hypothesis from different perspectives:

(i) For matrix decomposition methods, it is hard to promote the reconstruction accuracy when the CR gets larger, especially when the size of network gets larger. That is because the low-dimensional intra-day trend has been archived by the methods and the major errors remain in other components such as the burst components and Gaussian type fluctuation.

(ii) The proposed compression method significantly promotes the compression performance, the maximum of which reaches to 23.2%, especially when the CR is larger than 0.1. This finding directly validates our hypothesis.

(iii) Further analysis on the results show that the reconstruction error decrease changes very slowly and maintains at a relatively certain level as CR increases. According to our hypothesis that is because the low-dimensional intra-day trend and the burst components have been archived by the proposed compression method and the major errors may remain in the Gaussian type fluctuation.

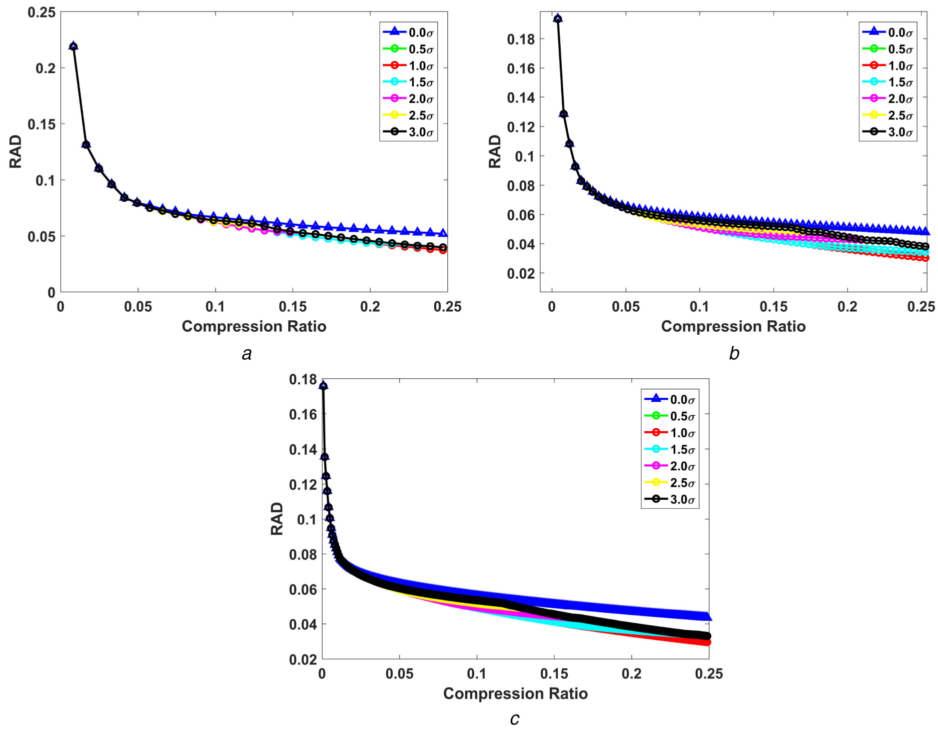

Fig. 11 *Sensitivity analysis of the threshold of burst points. The test network is* ***(a)*** CATC, ***(b)*** CATB, ***(c)*** CATC

Moreover, to implement the work technically, we compare the influence of different data reformatting, archive the bursts defined before in descending order with respect to the absolute value of the burst points and propose a flexible compression framework to balance between burst components and low-dimensional intra-day trend. The findings are as follows:

(i) For the fairness of comparison, we should take the local optimum format for different algorithms at different networks in the experiments. That is because there is no constantly best selection of data reformatting for all different sizes of traffic networks. In essence, each type of matrix mapping preserves some type of similarity but yields discrepancy simultaneously and it is the balance that really matters.
(ii) The policy we employ to archive the principal information of the bursts in this work is simple yet effective.
(iii) The proposed flexible compression framework in this work balances between burst components and low-dimensional intra-day trend well. It promotes the reconstruction accuracy significantly and provides more flexible choices for CR.

Our future work will focus on the following aspects. First, the fundamental idea of the hypothesis could be extended into other applications such as traffic flow prediction and missing data imputing. Second, the stochastic properties of the burst components and more effective methods to archive the principal information of the bursts are worthy of attention and effort. Besides intra-day trend and burst components, in order to further promote the compression performance, our team will study the fluctuation characteristics and archive the dominant information of the fluctuation.

6 Acknowledgments

This work was supported by the National Basic Research Program of China (973 Project) 2012CB725405, the national science and technology support programme (2014BAG03B01), the National Natural Science Foundation of China 51278280, the project supported by Tsinghua University (20131089307) and the Foundation of Beijing Key Laboratory for Cooperative Vehicle Infrastructure Systems and Safety Control.

7 References

[1] Zhang, J., Wang, F.Y., Wang, K., *et al.*: 'Data-driven intelligent transportation systems: a survey', *IEEE Transp. Intell. Syst.*, 2011, **12**, pp. 1624–1639
[2] Ke, R., Li, Z., Kim, S., *et al.*: 'Real-time bidirectional traffic flow parameter estimation from aerial videos', *IEEE Trans. Intell. Transp. Syst.*, 2017, **18**, (4), pp. 890–901
[3] Asif, M.T., Kannan, S., Dauwels, J., *et al.*: 'Data compression techniques for urban traffic data'. 2013 IEEE Symposium on Computational Intelligence in Vehicles and Transportation Systems (CIVTS), April 2013, pp. 44–49
[4] Leduc, G.: 'Road traffic data: collection methods and applications', *Working Papers Energy Transp. Clim. Change*, 2008, **1**, (55)
[5] 'Performance measurement system (PeMS)'. Available at http://pems.dot.ca.gov/, accessed February 2016
[6] Shawe-Taylor, J., De Bie, T., Cristianini, N.: 'Data mining, data fusion and information management', *Proc. Inst. Electr. Eng., Intell. Transp. Syst.*, 2006, **153**, (3), pp. 221–229
[7] Qu, L., Hu, J.M., Zhang, Y.: 'A flow volume data compression approach for traffic network based on principal component analysis'. Proc. 2007 IEEE Intelligent Transportation Systems Conf., Seattle, WA, USA, 2007, pp. 125–130
[8] Jun, D., Zuo, Z, Xiao, M.: 'A method for urban traffic data compression based on wavelet-PCA'. Fourth Int. Joint Conf. Computational Sciences and Optimization, 2011, pp. 1030–1034
[9] Shuo, F., Yi, Z., Li, L.: 'A comparison study for traffic flow data comparison'. 2016 12th World Congress on Intelligent Control and Automation (WCICA), Guilin, China, 12–15 June 2016, pp. 977–982
[10] Ou, X., Ren, J., Zhang, Y.: 'A neural network model for urban volumes compression'. The Sixth World Multicongress on Systemics, Cybernetics and Informations, Orlando, FL, June 2002
[11] Yang, X., Yu-ming, X., Lingyun, L., *et al.*: 'The traffic data compression and decompression for intelligent traffic systems based on two-dimensional wavelet transformation'. Seventh Int. Conf. Signal Processing, Beijing, China, 31 August–4 September 2004, pp. 2560–2563
[12] Steeb, W.H., Hardy, Y.: 'Matrix calculus and Kronecker product: a practical approach to linear and multilinear algebra', *World Sci.*, 2011
[13] De Lathauwer, L., De Moor, B., Vandewalle, J.: 'on the best rank-1 and rank-(r 1, r 2, …, rn) approximation of higher-order tensors', *SIAM J. Matrix Anal. Appl.*, 2000, **21**, (4), pp. 1324–1342
[14] Cichocki, A., Mandic, D., De Lathauwer, L., *et al.*: 'Tensor decompositions for signal processing applications', *IEEE Signal Process. Mag.*, 2015, **32**, (2), pp. 145–163
[15] Salomon, D.: '*Data compression: the complete reference*' (Springer Science & Business Media, 2004)
[16] Shannon, C.E.: 'A mathematical theory of communication', *Bell Syst. Tech. J.*, 1952, **27**, pp. 379–423, 623-656, July, October, 1948-1101
[17] Huffman, D.A.: 'A method for the construction of minimum-redundancy codes'. Proc. I.R.E., p. 1098
[18] Cover, T.M., Thomas, J.A.: '*Elements of information theory*' (Wiley-Interscience Press, 2006, 2nd edn.)
[19] Berger, T.: 'Rate distortion theory: mathematical basis for data compression', 1971

[20] Li, L., Su, X., Zhang, Y., *et al.*: 'Trend modeling for traffic time series analysis: an integrated study', *IEEE Trans. Intell. Transp. Syst.*, 2015, **16**, (6), pp. 3430–3439
[21] Chen, C., Wang, Y., Li, L., *et al.*: 'The retrieval of intra-day trend and its influence on traffic prediction', *Transp. Res. C*, 2012, **22**, pp. 103–118
[22] Li, L., Su, X., Wang, Y., *et al.*: 'Robust causal dependence mining in big data network and its application to traffic flow predictions', *Transp. Res. C*, 2015, **58**, pp. 292–307
[23] Li, L., Li, Z., Zhang, Y., *et al.*: 'A mixed-fractal traffic flow model whose Hurst exponent appears crossover'. Fifth Int. Joint Conf. Computational Sciences and Optimization, 2012, pp. 443–447
[24] Chen, Y., Li, L., Zhang, Y., *et al.*: 'Fluctuations in urban traffic networks', *Mod. Phys. Lett. B*, 2011, **22**, (2), pp. 101–115
[25] Eftekhari, H.R., Ghatee, M.: 'An inference engine for smartphones to preprocess data and detect stationary and transportation modes', *Transp. Res. C, Emerging Technol.*, 2016, **69**, pp. 313–327
[26] Asadi, R., Ghatee, M.: 'A rule-based decision support system in intelligent hazmat transportation system', *IEEE Trans. Intell. Transp. Syst.*, 2015, **16**, (5), pp. 2756–2764