Analysis of network-wide transit passenger flows based on principal component analysis

Luo, Ding; Cats, Oded; Van Lint, Hans

DOI
10.1109/MTITS.2017.8005611

Publication date
2017

Document Version
Accepted author manuscript

Published in
5th IEEE International Conference on Models and Technologies for Intelligent Transportation Systems, MT-ITS 2017 - Proceedings

Analysis of Network-wide Transit Passenger Flows Based on Principal Component Analysis

Ding Luo, Oded Cats and Hans van Lint
Department of Transport and Planning
Delft University of Technology, Delft, The Netherlands
Email: {d.luo, o.cats, j.w.c.vanlint}@tudelft.nl

***Abstract*—Transit networks are complex systems in which the passenger flow dynamics are difficult to capture and understand. While there is a growing ability to monitor and record travelers' behavior in the past decade, knowledge on network-wide passenger flows, which are essentially high-dimensional multivariate data, is still limited. This paper describes how Principal Component Analysis (PCA) can be leveraged to develop insight into such multivariate time series transformed from raw individual tapping records of smart card data. With a one-month data set of the Shenzhen metro system used in this study, it is shown that a great amount of variance contained in the original data can be effectively retained in lower-dimensional sub-spaces composed of top few Principal Components (PCs). Features of such low dimensionality, PCs and temporal stability of the flow structure are further examined in detail. The results and analysis provided in this paper make a contribution to the understanding of transit flow dynamics and can benefit multiple important applications for transit systems, such as passenger flow modeling and short-term prediction.**

***Keywords*—Transit system, smart card data, multivariate passenger flows, principal component analysis**

I. Introduction

Transit systems have been rapidly developed in many places as a solution to mobility and environmental problems, especially in metropolitan areas with dense population and limited road resources. These transit systems, mostly comprised of lines of buses, trams and metros, are being used by a great number of travelers every day for all kinds of activities. As many transit systems are still expanding and attracting more travelers, it becomes imperative for transit researchers, managers and operators to gain more knowledge on such complex systems, which can largely benefit the development of advanced transit management tools. For example, one of the significant tasks is to understand the dynamics of passenger flows in transit systems in order to facilitate advanced transit fleet and demand management. This particular task, which was until recently hindered by limited amount of flow observations, has now been enabled thanks to the vast amount of smart card data collected by automatic fare collection (AFC) systems [1]. Since individual travelers' trips are passively recorded while they travel, the information contained in such data is very complete and characterized by fine granularity. It hence provides both practitioners and researchers with a precious chance to investigate transit mobility and flow dynamics in depth.

Numerous studies which leverage smart cart data to unravel the spatial-temporal patterns of transit trips, urban mobility and travelers' behaviors have been published, such as [2]. Besides, researchers have also attempted to measure the variability of mobility patterns [3] and identify urban activity centers or clusters [4] based on passenger flows obtained from smart card data. These existing studies succeeded in strengthening our understanding of urban mobility and transit systems, but a limited number of them were found to shed light upon passenger flow dynamics from a multivariate perspective, which means to deal with the high dimensionality of such flow data. Due to the existence of the so-called "curse of dimensionality" [5], the development of a series of important transit applications, such as network-wide flow modeling and prediction, might be hindered without sufficient insight into these multivariate passenger flows. It becomes substantially difficult to find effective and intuitive solutions in a high-dimensional space while dealing with complex systems like a transit network which consists of multiple lines and hundreds of stations. As argued by [5], this particular difficulty results from the conjunction of two effects. Firstly, some geometrical properties of high-dimensional spaces are counter-intuitive and different from what can be observed in 2- or 3-dimensional spaces. Secondly, data analysis tools are usually designed in low-dimensional spaces with intuition.

Given the research gap and difficulty described above, this study is aimed at performing a multivariate analysis of transit passenger flows based on a well-known dimensionality reduction technique, Principal Component Analysis (PCA). We detail how a one-month smart cart data set from the Shenzhen metro system is transformed to multivariate time series of flows and how PCA is performed on such time series. The results of PCA, including the low dimensionality of flows, features of principal components (PCs), approximation of original flows, and temporal stability of flow structure, are explicitly presented and analyzed, providing an insight into the underlying structure of flow dynamics within a complex transit network. Overall, this study contributes to the development of multivariate analysis on transit passenger flows, and shows the potential of incorporating PCA into promising applications, such as anomaly detection and short-term forecasting.

The remainder of this paper is organized as follows. Section II gives a general background of PCA and the specific procedure of applying PCA to multivariate time series. Then a detailed description of transit network and data studied is provided in section III, followed by the presentation of results and analysis. Conclusion is drawn in section V with suggestions for future research directions.

II. Principal Component Analysis

A. Background

PCA was initially proposed to describe the variation of a set of uncorrelated variables in a multivariate data set [6], [7]. So far it has been extensively used as a technique to perform various tasks, such as dimensionality reduction, factor analysis, feature extraction, and lossy data compression. In the field of traffic and transportation, for example, PCA was utilized to compress traffic network flow data [8], and was integrated into dynamic origin-destination (O-D) estimation and prediction in order to overcome the computational problem caused by high-dimensional O-D matrix data [9]. It was also leveraged to analyze travelers' longitudinal behavior by extracting the so-called eigen-sequences [10], to name a few.

The central idea of PCA is to reduce the dimensionality of a data set consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set [11]. PCA achieves this target by projecting the observations onto a new set of axes which are called the PCs. Each PC has the property that it points in the direction of maximum variance remaining in the data, given the variance already accounted for in the preceding components. As such, the first PC captures the total energy of the original data to the maximal degree possible on a single axis. The following PCs then capture the maximum residual energy among the remaining orthogonal directions. In this sense, the PCs are ordered by the amount of energy in the data they capture.

B. Application to Flow Analysis

It has been shown that PCA can be used as an effective tool to analyze whole-network traffic flows which are essentially high-dimensional multivariate time series [12]. By performing PCA on the flow data, a smaller number of dimensions can be found and leveraged to well approximate original high-dimensional data. Let $\mathbf{X}$ denote a matrix of multivariate flow time series as equation (1) shows. Each column i of $\mathbf{X}$ denotes a single flow variable, while each row j represents an observation of all flow variables at time j. This yields a $t \times p$ matrix $\mathbf{X}$, where t represents the total number of time instances and p represents the total number of flow variables.

$$\mathbf{X} = \begin{bmatrix} x_1(1) & x_2(1) & \dots & x_p(1) \\ x_1(2) & x_2(2) & \dots & x_p(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t) & x_2(t) & \dots & x_p(t) \end{bmatrix} \tag{1}$$

As shown in equation (2), obtaining all the PCs of $\mathbf{X}$ is actually equivalent to calculating the eigenvectors of $\mathbf{X^T X}$ which is a measure of the covariance between flows [11].

$$\mathbf{X^T X v_i} = \lambda_i \mathbf{v_i} \tag{2}$$

where λ_i is the eigenvalue corresponding to eigenvector $\mathbf{v_i}$ $(p \times 1)$ and the number of eigenvalues/eigenvectors is equal to the number of variables p. In fact, the eigenvalue λ_i indicates how much variance of the original data is explained by the dimension i specified by eigenvector $\mathbf{v_i}$.

$$\text{Var}(\mathbf{v_i}^\mathbf{T} \mathbf{X}) = \lambda_i \tag{3}$$

Arranging all the eigenvalues in a descending order ($\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_p$), the first PC is thus the eigenvector which corresponds to the largest eigenvalue since it accounts for the greatest variance in the entire data.

By mapping the original data onto the derived principal component space, it can be seen that the contribution of dimension i (the i-th PC) as a function of time is given by $\mathbf{X v_i}$. Normalizing this vector to unit length with λ_i as shown in equation (4), we obtain a $t \times 1$ vector $\mathbf{u_i}$ which contains the information of temporal variation along the i-th PC. As a matter of fact, the vector $\mathbf{u_i}$ captures the temporal variation common to all flows along this dimension (PC). The set of vectors $[\mathbf{u_1}, \mathbf{u_2}, \dots, \mathbf{u_p}]$, which are perpendicular, can thus be referred to as the **eigen-flows** of $\mathbf{X}$.

$$\mathbf{u_i} = \frac{\mathbf{X v_i}}{\sqrt{\lambda_i}} \tag{4}$$

Let $\mathbf{V}$ denote a $p \times p$ matrix consisting of all the PCs $[\mathbf{v_1}, \mathbf{v_2}, \dots, \mathbf{v_p}]$ which are arranged in order. The first column $\mathbf{v_1}$ refers to the first PC, and so on. Let $\mathbf{U}$ denote a $t \times p$ matrix of which column i is $\mathbf{u_i}$. Consequently, each individual flow $\mathbf{X_i}$ can be written as:

$$\frac{\mathbf{X_i}}{\sqrt{\lambda_i}} = \mathbf{U}(\mathbf{V^T})_\mathbf{i} \tag{5}$$

where $\mathbf{X_i}$ is the time series of i-th flow and $(\mathbf{V^T})_\mathbf{i}$ is the i-th row of $\mathbf{V}$. This equation indicates that each flow $\mathbf{X_i}$ is essentially a linear combination of the eigen-flows with weights specified by $(\mathbf{V^T})_\mathbf{i}$.

By selecting the first r $(r \leq p)$ eigenvectors with largest eigenvalues, the information contained in original data $\mathbf{X}$ can then be effectively transformed onto a r-dimensional subspace of $\mathbb{R}^p$. It is shown in equation (6) how the approximation can be done.

$$\mathbf{X'} \approx \sum_{i=1}^{r} \sqrt{\lambda_i} \mathbf{u_i} \mathbf{v_i^T} \tag{6}$$

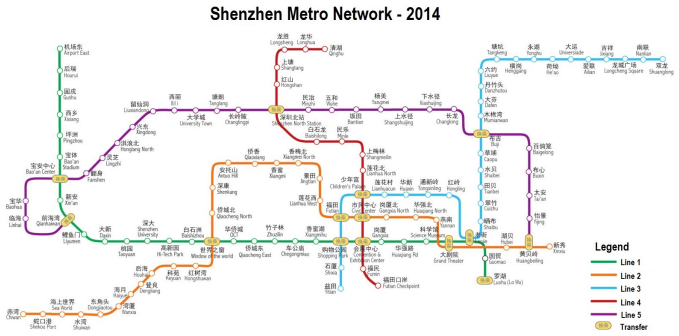

Fig. 1. An illustration of Shenzhen metro network (2014).

III. Transit Network and Data

A. Shenzhen Metro Network

The network of Shenzhen metro system by the end of 2014 was studied. Shenzhen is one of the largest cities in China with a metropolitan area population of over 18 million. By the end of 2014, there were in total five lines in operation with 118 stations. An illustration of the network is shown in Fig. 1, where five lines are represented with different colors, and all transfer stations are highlighted with yellow marks. As a major transportation service in Shenzhen, the metro system accounts for approximately one third of the total public transport passenger traffic, which results in complex passenger flow dynamics over time. It is therefore a significant task to understand these flow dynamics for achieving better system operations and management.

B. Smart Card Data

AFC system was employed by the Shenzhen metro system and passengers could not travel without using a smart card. Moreover, tapping is required for both entry and exit activities because the fare was collected using a distance-based scheme. As a result, complete travel information of individuals except for transfer activities were recorded in the database. A typical record includes the time-invariant anonymous card ID, metro station ID, transaction timestamp, and transaction type (21 for tap-in and 22 for tap-out). The data set used for this study contains 139,646,884 records, covering the whole period of December of 2014. The period includes 23 normal weekdays and four weekends.

C. Entry and Exit Flow Profiles

Entry and exit flows at stations were investigated in this study. Raw data were transformed from a per-user basis to a per-station basis with time discretization. Time series of both entry and exit flows for each station with a time interval of 5 minutes were constructed. In line with the operational time of the metro system, the time horizon considered in this study was from 6 AM to 11 PM each day (17 hours). The total number of measurements for each flow variable over the entire period is then 6324 ($= 12 \times 17 \times 31$). Cumulative distribution function (CDF) plots of these flow profiles are shown in Fig. 2a. It can

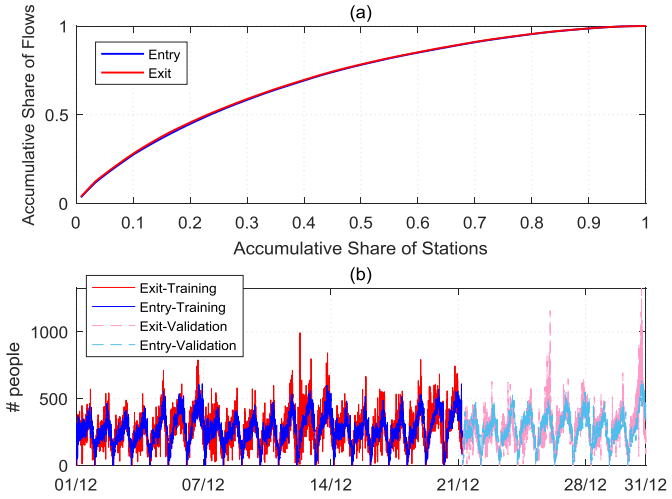

Fig. 2. Illustrations of flow profiles. (a) Cumulative distribution function plots of entry and exit flows; (b) A typical example of entry and exit flow time series of Shenzhen metro station (Luohu station).

be understood from the diagrams that, in this case, flows are accounted for by a relatively large percentage of stations rather than only a few. This feature will be reflected in the following analysis.

D. Training and Validation Sets

In order to examine whether the structure of entry and exit flows is temporally stable, the entire data set was divided into training part and validation part, with the former containing the data from the first three weeks (2014-12-01 to 2014-12-21) and the latter from the rest of the days (2014-12-22 to 2014-12-31). Note that Christmas day and New Year Eve are included in the validation set. The main motivation is to know whether the decomposition of entry and exit flows into eigen-flows, as determined by the set of PCs, is useful for analyzing data that are not part of the input to the PCA procedure. This is crucial for applications like forecasting. By using the training data alone, we obtained a 4284×236 flow matrix following equation (1). With p equal to 236, the first half of the columns were filled with entry flows while the second half with exit flows. An advantage of using PCA is that both entry and exit flows can be analyzed simultaneously, thus allowing us to obtain insight into underlying patterns of the network-wide flows.

IV. Results and Analysis

A. Low Dimensionality of Flows

PCA was performed on the training flow matrix which was specified in the previous section. Since the magnitude of eigenvalues indicates how much variance is explained by the corresponding eigenvector, which is equivalent to PC, a scree plot shown in Fig. 3a based on eigenvalues can be leveraged to conduct visual examination. It can be seen through the sharp knee of the curve that the majority of variance contained in the data is virtually contributed by the first few eigen-flows,

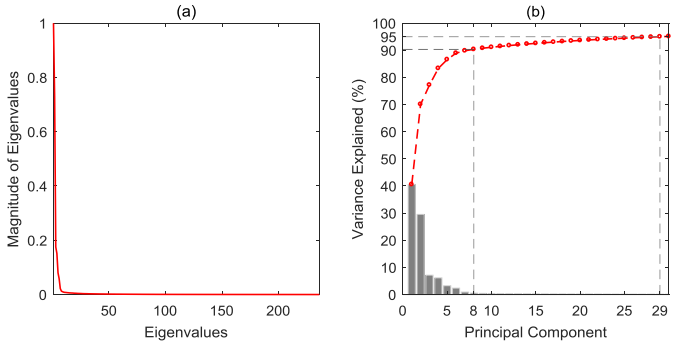

Fig. 3. Demonstration of the low dimensionality of entry and exit flows. (a) Scree plot of eigenvalues; (b) Cumulative percentage of the total variance explained by PCs (principal components). Over 90% variance can be explained by only 8 PCs, while over 95% can be explained by 29 PCs.

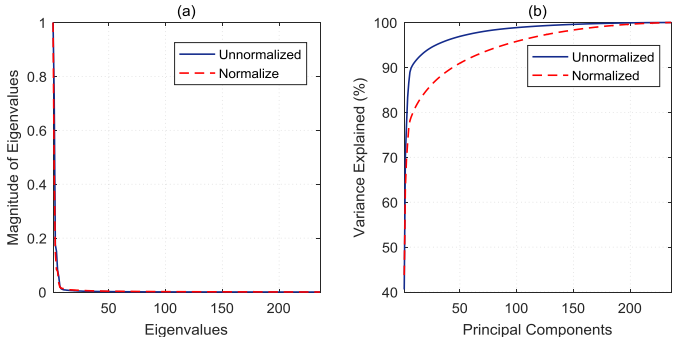

Fig. 4. Comparison of PCA results for normalized and unnormalized flows. (a) Scree plots based on eigenvalues; (b) Cumulative distribution function plots.

namely the temporal variability on the first few PCs. Fig. 3b further explicitly displays that 8 and 29 PCs, respectively, can account for over 90% and over 95% variance in the data.

There are two possible explanations to such intrinsic low dimensionality of multivariate flow time series. The first one is that it may be attributed to the fact that variation along a small set of dimensions in the original data is dominant. The second reason is that non-negligible correlation among variables may matter greatly, which implies the common underlying patterns or trends across dimensions. In order to understand how each of these two factors accounts for the variance in the current case, PCA can be as well performed on normalized flow variables with zero mean and unit variance. The normalization is specified by equation (7). The motivation is that if the low dimensionality still exists after normalization, it can be concluded that the correlation among flows plays the most important role because the normalization procedure has already removed the effect of magnitude in all original flows.

$$\bar{\mathbf{X}}_i = \frac{\mathbf{X}_i - \mu_i}{\sigma_i} \tag{7}$$

where μ_i and σ_i denote the sample mean and variance of the i-th column of $\mathbf{X}$.

A comparison between normalized and unnormalized cases is illustrated in Fig. 4. No striking difference between two scree plot curves can be seen in Fig. 4a, which indicates that the vast majority of low dimensionality of these flows is actually a result of the correlations among them. Moreover, Fig. 4b displays that the difference in flow magnitude also accounts for the low dimensionality to some limited extent because more PCs are needed to retain as much variance as before. This important finding coincides with the nature of Shenzhen metro system reflected by Fig. 2a in which the CDF curve does not quickly reach a relatively flat level. Such pattern indicates that this is not the case where only a few flows are completely dominant. How these PCs relate to flow and demand patterns are further illuminated in the following section.

B. *Principal Components and Eigen-flows*

Three typical examples of PCs (236 in total) and corresponding eigen-flows are demonstrated in Fig. 5. While the top eigen-flow evidently shows weekly periodicity, the other two mostly show randomness with the middle one having two noticeable spikes. Clearly, the first PC very well captures the morning and afternoon peaks of passenger flows in the metro system. The spikes in the middle plot, however, are normally a sign of some special occurrences in the data. Following the taxonomy proposed by [12], the top, middle, and bottom eigen-flows can be roughly referred to as the deterministic, spike, and noise ones, respectively, though in the current case the spike eigen-flows are not always sufficiently significant. This is mainly because there are not many irregular observations of flows in the training set.

The so-called eigen-flows essentially capture all original flows' temporal variation projected onto the PCs. These PCs, shown in the right column of Fig. 5, specifically determine how their corresponding eigen-flows contribute to each original flow. Therefore the matrix $\mathbf{V}$ consisting of PCs is also called a *loading* matrix or coefficients. It can be observed that top eigen-flows (variability on top PCs) make greater contribution to original flows. This is consistent with the low dimensionality of original flows.

It can be further investigated how many eigen-flows significantly contribute to one single original flow. This can be done by checking whether a loading coefficient on that row is larger than $\sqrt{p}$. With p equal to 236 in this case, such threshold would be 0.0651. This standard is deemed reasonable because a perfectly equal mixture of all eigen-flows would result in a row of $\mathbf{V}$ with all entries equal, under the condition that columns of $\mathbf{V}$ have unit norm. The result of applying this rule to all rows of $\mathbf{V}$ is illustrated through a CDF plot in Fig. 6a. It shows none of the original flows needs more than 70 significant PCs for sufficient reconstruction. In fact, about half of them are composed of less than 45 significant PCs, implying that each entry or exit flow only possesses a relatively small set of temporal variability features.

Furthermore, it is also possible to understand how the

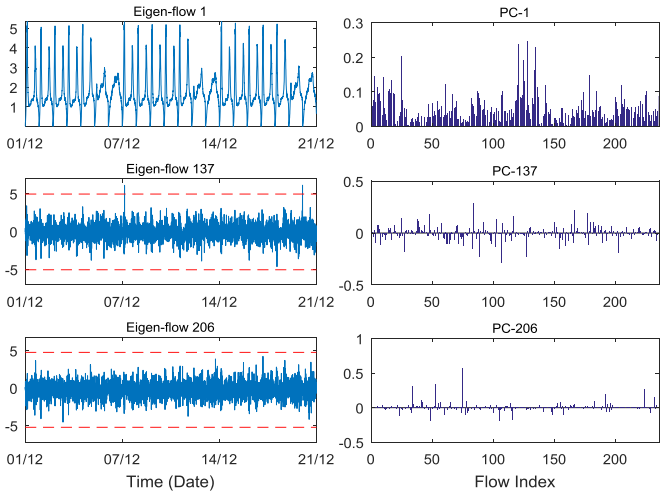

Fig. 5. Illustration for examples of eigen-flows and PCs.

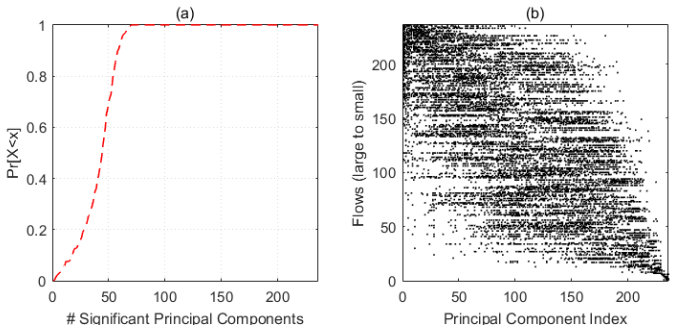

Fig. 6. Illustration of analysis on flow structure. (a) CDF plot of the number of significant PCs needed for original flows; (b) A scatter plot showing how every single flow is significantly contributed by PCs. The flow index is arranged from top to bottom in a descending order in terms of flow magnitude, while the PC index is arranged from left to right in a descending order in terms of the variance it explains.

magnitude of a single flow is linked to the significant PCs (loading coefficient larger than $\sqrt{p}$) that constitute it. A graph shown in Fig. 6b is used for better visual examination. Its horizontal axis represents the index of PCs ordered in a descending sequence in terms of their eigenvalues (how much variance they explain), and the vertical axis represents the index of original entry and exit flows which are also ordered in a descending sequence in terms of their average magnitude. Once a loading coefficient is larger than $\sqrt{p}$ in absolute value, a black dot will be plotted at that spot. In this manner, the top rows in the graph demonstrate the PCs that are significant in explaining the temporal variability of strong flows, while the bottom rows show those that are significant for weak flows. Although all the dots in Fig. 6b scatter considerably, a general diagonal trend from upper left to lower right can be identified. It implies that larger flows tend to be composed of most significant PCs, and vice visa (smaller flows tend to be composed of insignificant PCs). This feature pertains to the approximation of original flows using PCs, which is going to be discussed in the following section.

C. Reconstruction of Original Flows using PCs

According to equation (6), the original flows can be approximated using a set of selected PCs. Essentially, such approximation is realized by forming a linear combination of eigen-flows. Fig. 7 demonstrates three typical examples of reconstructed flow time series using both 8 (90% total variance explained) and 29 PCs (95% total variance explained). Specifically, the left column displays the overall time series for three weeks, while the right one zooms into a more detailed level with only one day illustrated.

As the busiest metro station in Shenzhen, Laojie station's entry flow profile (top) can be well reconstructed in both scenarios. The middle row, however, shows a different case where the original data (Luohu entry flow) can be approximated much better with 29 PCs than 8 PCs. Note Luohu is a metro station heavily used by travelers because of the train station and port connecting Shenzhen and Hong Kong next to it. The bottom row then shows something different from both above. It can be seen that the entry flow at AntuoHill station is very low all the time, and both approximations cannot capture detailed fluctuations but the main trend of the time series. These results are in line with the low dimensionality of flows and the implication obtained from the last section that larger flows tend to be composed of most significant PCs, while smaller ones tend to be composed of insignificant PCs.

D. Temporal Stability of Flow Structure

With the whole data set divided into training and validation parts, it can be examined whether the PCs derived from the previous time period can be also used to approximate future time series. It is valuable to know if the underlying flow patterns are stable over time so that this intrinsic feature can benefit multiple applications, such as anomaly detection and short-term prediction. In practice, the PCs contained in the loading matrix $\mathbf{V}$ was computed using the training set, and then these PCs were used to derive eigen-flows of the validation data set based on equation (4). Eventually, the same approximation procedure can be performed using equation (6), of which results are shown in Fig. 8.

Overall, it can be observed that the PCs computed using past data are still capable of approximating the profile of future data. Nevertheless, it should be noticed that the effectiveness of these "old" PCs can be diminished when non-recurrent or special events occur, as shown in the top row of Fig. 8 which demonstrates the entry flows of Laojie station on the new year eve (December 31, 2014). With more incoming passengers than an average weekday, it can be seen that the approximation results of both cases (8 and 29 PCs) illustrate more deviation from the original flow as well. The comparison between the top right plots of Fig. 7 and Fig. 8 makes it even clearer. As to the other two examples, however, the distinction is not as clear as the case of Laojie station. It can thus be concluded that the structure of flows captured by PCs can still be temporally

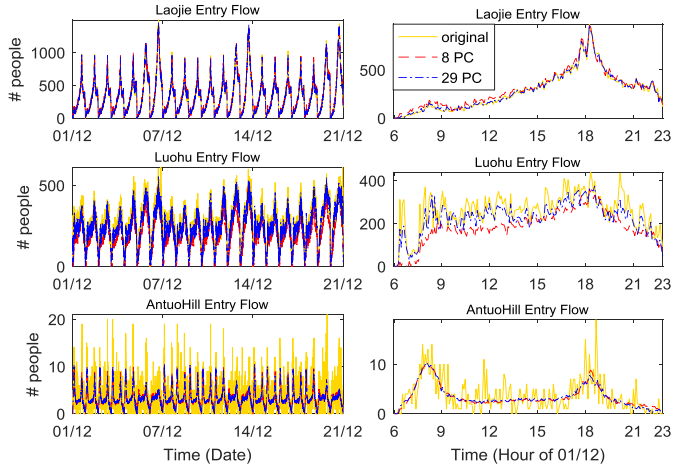

Fig. 7. Examples of approximating original flows using different number of PCs. The left column illustrates the results of the entire period covered by the training data, while the right column shows the zoom-in plots of the first day (December 1, 2014).

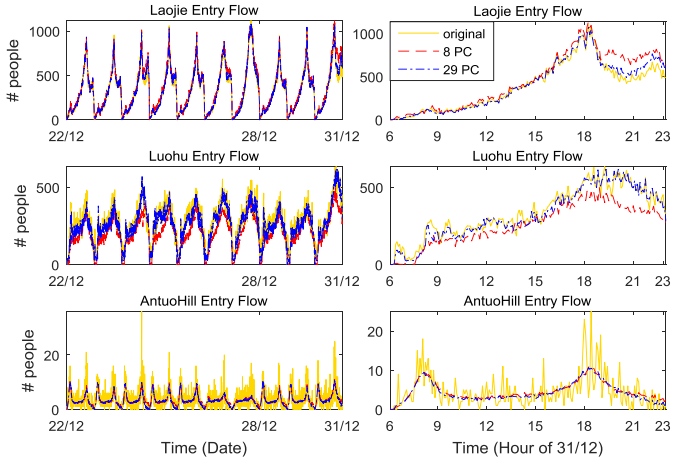

Fig. 8. Examples of approximating flows using PCs that are not computed based on these flow data. The left column illustrates the results of the entire period covered by the validation data, while the right column shows the zoom-in plots of the last day (December 31, 2014).

stable, mostly for normal days, but failure can occur when there are uncommon changes in flows.

V. Conclusion

This paper demonstrates how Principal Component Analysis (PCA) can be applied to multivariate transit passenger flows as a solution to high-dimensional data analysis problems. With a one-month smart card data set from the Shenzhen metro system leveraged, PCA is performed on a 4284×236 multivariate time series matrix (entry and exit flows considered simultaneously) transformed from the original individual tapping records. The results and analysis show that a great amount of variance contained in the original data can be effectively retained in lower-dimensional sub-spaces composed of top few Principal Components (PCs). This feature of low dimensionality is thus thoroughly examined in the study, with the essence of PCs and eigen-flows, as well as the temporal stability of PCs revealed in the subsequent investigation.

As the availability of advanced transit data increases rapidly, new chances of understanding the dynamics of complex transit systems have been opened up. Future studies can explore how the low dimensionality of flows can be incorporated in transit flow modeling and prediction. Moreover, other non-linear dimension reduction techniques which can keep the nonlinear relations between the components of the initial data, as mentioned in [5], can also be tested and compared in the context of transit systems.

Acknowledgment

The authors would like to acknowledge the support of the SETA project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No. 688082.

References

[1] M.-P. Pelletier, M. Trépanier, and C. Morency, "Smart card data use in public transit: A literature review," *Transportation Research Part C: Emerging Technologies*, vol. 19, no. 4, pp. 557–568, 2011.

[2] L. Sun, K. W. Axhausen, D.-H. Lee, and X. Huang, "Understanding metropolitan patterns of daily encounters," *Proceedings of the National Academy of Sciences*, vol. 110, no. 34, pp. 13 774–13 779, 2013.

[3] C. Zhong, E. Manley, S. M. Arisona, M. Batty, and G. Schmitt, "Measuring variability of mobility patterns from multiday smart-card data," *Journal of Computational Science*, vol. 9, pp. 125–130, 2015.

[4] O. Cats, Q. Wang, and Y. Zhao, "Identification and classification of public transport activity centres in stockholm using passenger flows data," *Journal of Transport Geography*, vol. 48, pp. 10–22, 2015.

[5] M. Verleysen and D. François, *The Curse of Dimensionality in Data Mining and Time Series Prediction*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 758–770.

[6] K. Pearson, "Liii. on lines and planes of closest fit to systems of points in space," *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, vol. 2, no. 11, pp. 559–572, 1901.

[7] H. Hotelling, "Analysis of a complex of statistical variables into principal components." *Journal of educational psychology*, vol. 24, no. 6, p. 417, 1933.

[8] Q. Li, H. Jianming, and Z. Yi, "A flow volumes data compression approach for traffic network based on principal component analysis," in *Intelligent Transportation Systems Conference, 2007. ITSC 2007. IEEE*. IEEE, 2007, pp. 125–130.

[9] T. Djukic, J. Van Lint, and S. Hoogendoorn, "Application of principal component analysis to predict dynamic origin-destination matrices," *Transportation Research Record: Journal of the Transportation Research Board*, no. 2283, pp. 81–89, 2012.

[10] G. G. Langlois, H. N. Koutsopoulos, and J. Zhao, "Inferring patterns in the multi-week activity sequences of public transport users," *Transportation Research Part C: Emerging Technologies*, vol. 64, pp. 1–16, 2016.

[11] I. Jolliffe, *Principal component analysis*, 2nd ed. Springer-Verlag New York, 2002.

[12] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft, "Structural analysis of network traffic flows," *SIGMETRICS Perform. Eval. Rev.*, vol. 32, no. 1, pp. 61–72, Jun. 2004.