Measuring Variability in Urban Traffic Flow by Use of Principal Component Analysis

THEODORE TSEKERIS
ANTONY STATHOPOULOS*

Department of Transportation
Planning and Engineering
School of Civil Engineering
National Technical University of Athens
5 Iroon Polytechniou
157 73 Athens, Greece

ABSTRACT

This paper presents a new approach for the spatio-temporal analysis of variation in traffic flow. Traffic detectors located in several arterial links of an extended urban network yield the time series of aggregate data used in the approach, which is based on the Principal Component Analysis (PCA) of these time series spanning several weeks. The analysis demonstrates the small variability in traffic flow over the whole network. The statistical analysis of common sources of temporal variation in traffic flow provides considerable insight into the properties of long-term flow dynamics. The approach was found to be capable of identifying the location and the impact of extreme events in the network.

INTRODUCTION

The increasing availability of traffic flow information from archived and readily available sources, such as inductive loop detectors, prompts the ongoing development of data fusion and processing techniques for the fast and efficient analysis of network congestion problems. A major issue in tackling such problems is the measurement of the spatial and temporal variations in traffic flow. These variations are exceedingly useful as input to a wide variety of

Email addresses:
*Corresponding author—astath@transport.ntua.gr
T. Tserkeris—fabtse@central.ntua.gr

KEYWORDS: Traffic flow variability, urban networks, principal component analysis, smoothing models.

applications. Such applications include advanced systems that provide traffic information to travelers, the identification of erroneous traffic forecasts and extreme events (outliers) such as incident detection, the validation of traffic simulation models, and network capacity planning. Other applications refer to the design and evaluation of traffic management strategies, including traffic control and pricing policies, and the assessment of their environmental effects.

Nonetheless, existing applications of methods used for measuring traffic variability are mostly focused on freeways (Rakha and Van Aerde 1995), and they typically refer to a short temporal scale of analysis, ranging from a few seconds to several hours (Treiber and Helbing 2002). Moreover, the usage of correction factors to measure daily or monthly traffic variations based on annual average daily traffic (AADT) estimates (Sharma et al. 1996; Davis 1997), as obtained from traffic counts of medium-time period (usually of 24-hour period), cannot provide an indepth explanation of the sources contributing to the variability in urban traffic.

The investigation of traffic variability in urban arterial networks over long periods of analysis, spanning several weeks or months, can provide promising insight to the potential of the aforementioned applications to alleviate increasing congestion problems. Stathopoulos and Karlaftis (2001) first examined the spatio-temporal variations of traffic flow in a real urban network, the road network of the Greater Athens Area (GAA), Greece, by presenting an exploratory analysis of the distribution characteristics of a set of traffic measurements collected over a period of several months. Also, Weijermars and van Berkum (2004) presented an analysis of variance (ANOVA) of traffic flow along an urban route across a series of weekdays, based on the assumption that flows follow a normal distribution.

This paper describes a novel, interpretive approach for the simultaneous modeling of network-wide traffic flow time series collected over a one-month period from traffic detectors located at major arterial links. The approach, which is based on the general theory of linear algebra, explicitly recognizes the fact that some of these time series are both temporally and spatially correlated in the network, without relying on any a priori assumption concerning the distribution of traffic flows. More specifically, the method of Principal Component Analysis (PCA), also known as Singular Value Decomposition (SVD) (Meyer 2000), is applied in order to disentangle the intricate sources of long-term traffic dynamics manifested in large-scale urban networks, such as the GAA network.

This is achieved by identifying common underlying sources of temporal variability in traffic flow, which are obtained by estimating the eigenflows, originally defined in (Lakhina et al. 2004) to describe variations in origin-destination (OD) flows of Internet networks. An eigenflow is a time series that captures a common pattern (or source) of temporal variability in traffic flow at the network level. Each traffic flow time series is expressed as a weighted sum of eigenflows and the corresponding weights reflect the extent to which each source of temporal variability is present in the given traffic flow. The method of PCA in the context of traffic flows is analytically described in the second section.

The third section presents the traffic detector data used for the purposes of analysis. The fourth section describes how PCA can be employed to measure the variability of individual traffic flows and of aggregate network traffic, and implications of this measurement for traffic data reconstruction and traffic flow prediction. The fifth section provides a method for decomposing eigenflows to identify different sources of variability in traffic flow. Applications of this method for traffic modeling and incident detection in extended urban networks are also reported. The final section concludes the findings of the study.

METHOD OF PRINCIPAL COMPONENT ANALYSIS

The method of PCA provides the transformation (or mapping) of a dataset onto a new set of principal axes or components. These axes are ordered by the amount of variation (or energy) that they capture in the data. Namely, the first principal axis captures the maximum amount of variation that is possible to represent on a single axis. Each of the remaining

principal axes captures sequentially the maximum residual variation not captured by the preceding axes. In this way, the PCA offers a powerful tool for analyzing the total traffic variability in an urban-scale network composed of a large number of dimensions by approximating it within a lower dimensional structure that preserves its important properties.

Let m be the number of traffic detectors located on a subset of the total set of arterial links of an urban network, t be the number of successive days (e.g., the respective periods) in which the detector data are collected, and τ be the number of time intervals wherein each day is partitioned. The present study refers to realistic large-scale networks composed of thousands of links servicing hundreds of thousands of travelers. Such networks typically involve hundreds of detectorized links with traffic detector data aggregated over small time intervals, such as 15 minutes. Then, a matrix X can be defined, referred to here as measurement matrix, with p rows and m columns, where $p = \tau \times t$. Therefore, each column i of matrix X denotes the i-th traffic flow time series, represented by the column vector X_i, and each row j denotes the particular point in the time series in which traffic flows have been collected at interval j.

The calculation of the i-th principal component, v_i, is carried out through the spectral decomposition of the matrix $X^T X$, which provides a measure of the covariance between traffic flows, as follows:

$$X^T X v_i = \lambda_i v_i, \qquad i = 1,...,m \tag{1}$$

where λ_i is the non-negative real scalar, known as the eigenvalue, corresponding to principal component v_i. By convention, the eigenvalues are arranged in order of magnitude, from large to small, so that $\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_m$. By solving equation (1), the maximum variation of the measurement matrix X is captured by the first principal axis v_i. Proceeding recursively, once the first i-1 principal components have been determined, the i-th principal component corresponds to the maximum variation of the residual, that is the difference between the original data and the data mapped onto the first i-1 principal axes. The arrangement of the set of principal components $\{v_i\}_{i=1}^{m}$ in such an order, as columns, results in the principal matrix V, which has size $m \times m$. By definition, the columns of V matrix have unit norm, which means that the length of each column vector, as defined by the square root of the sum of squares of all entry values, is equal to unity.

Because the principal axes are arranged in order of contribution to the overall variation, the time-varying trend common to all flows along principal axis i can be represented through a column vector u_i with size p, referred to as the eigenflow of the i-th principal axis, as follows:

$$u_i = \frac{X_{v_i}}{\sigma_i}, \qquad i = 1,...,m \tag{2}$$

where $\sigma_i = \sqrt{\lambda_i}$ is the singular value corresponding to the i-th principal axis. The magnitude of singular values demonstrates the overall variation attributable to each particular principal component and, hence, the potential to reconstruct total traffic data using a smaller number of dimensions. The arrangement of the set of eigenflows $\{u_i\}_{i=1}^{m}$ as columns in order of decreasing strength of the common temporal trends results in the eigenflow matrix U, which has size $p \times m$. Based on equation (2), it can be shown that each traffic flow time series X_i, when normalized by the singular value σ_i, is a linear combination of the eigenflows, weighted by the associated principal component. More specifically, the relationship between matrices X, U, and V can be represented as follows:

$$\frac{X_i}{\sigma_i} = U(V^T)_i, \qquad i = 1,...,m \tag{3}$$

where $(V^T)_i$ is the i-th row of matrix V. By assuming that only a small number of $q < m$ singular values is non-negligible (see 2.2), or, in other words, only a small set of q eigenflows contributes to the bulk of temporal variability in traffic flow, then, the original data, that is, measurement matrix X, can be approximated as follows:

$$X' \approx \sum_{i=1}^{q} \sigma_i u_i v_i^T \tag{4}$$

The spatio-temporal reconstruction of matrix X by use of a lower number of dimensions can enhance the interpretability of the long-term dynamics of

each traffic flow, particularly for the case of large-scale urban networks, as shown in the fourth section.

TRAFFIC DETECTOR DATA

The traffic data used here for analysis purposes are automatically collected using loop detectors at 140 key locations around the urban road network of the Greater Athens Area (GAA), as illustrated in figure 1. These real-time data are stored at the end of every 90-sec signalization cycle and aggregated at time intervals of 15 minute duration. The traffic counting system provides an appropriate data quality control by performing screening and data repair functions so as to identify and exclude or smooth data from malfunctioning detectors. The dataset includes measurements corresponding to the first 28 of the 29 days in February 2000. Each day covers 16 time intervals of the period spanning between 6:00 am and 10:00 am. This yields a total number of 62,720 measurements, that is, a time series of 448 measurements for each of the 140 detector locations.

USE OF PCA FOR MEASURING TRAFFIC FLOW VARIABILITY

Spatio-Temporal Representation of Traffic Variations

Figure 2 presents a typical example of an eigenflow u_i (i=3) and its corresponding principal axis v_i, as calculated with the PCA of the four-week traffic detector data of the Greater Athens Area. Figure 2(a) demonstrates the representation of a pattern of temporal variation common to all traffic flow time series through eigenflow 3. The fifth section provides a systematic way for distinguishing different types of this temporal variation. Figure 2(b) illustrates the extent to which this particular temporal pattern is present in each traffic flow, through the entries of the corresponding principal component. Eigenflow 3 is most strongly present at traffic flow measurement point number 38, as shown in figure 2(b), namely, in the time series measured at detector number 38, whose location in the network is marked on figure 1. The immediately next strongest

FIGURE 1 Illustration of the Greater Athens Area (GAA) Network and Configuration of the Location of Loop Detectors

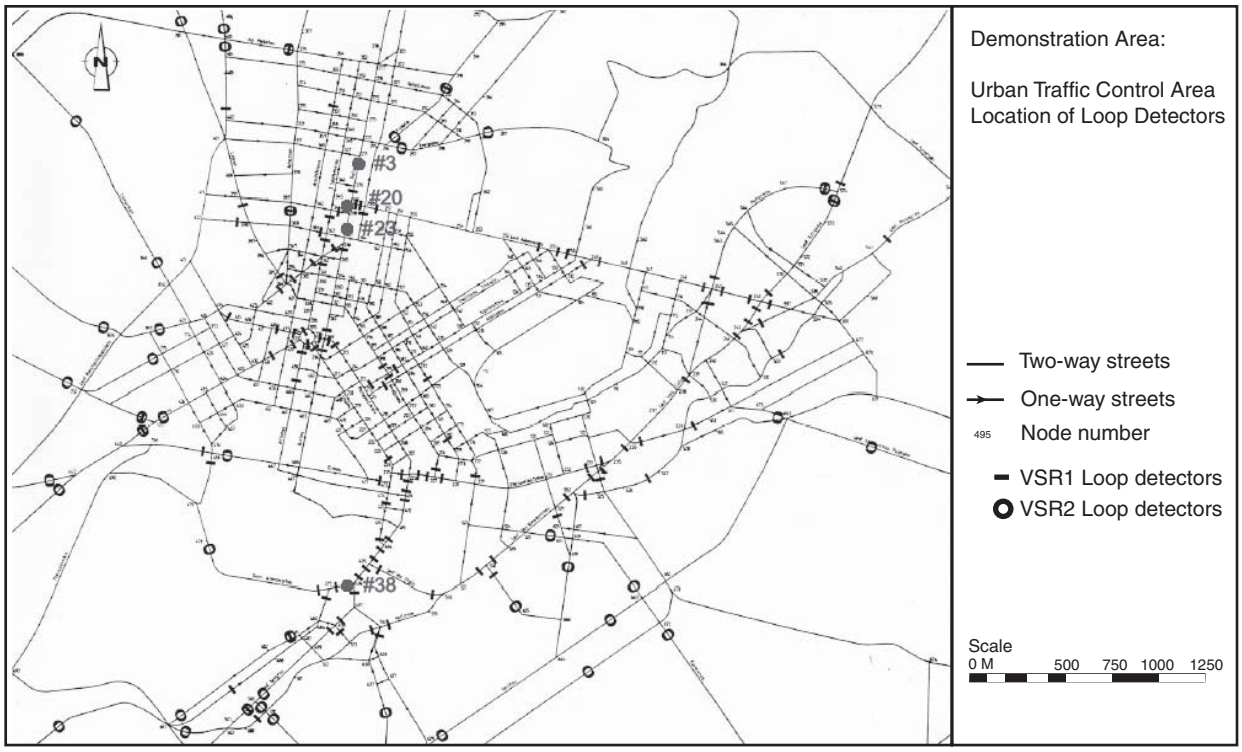
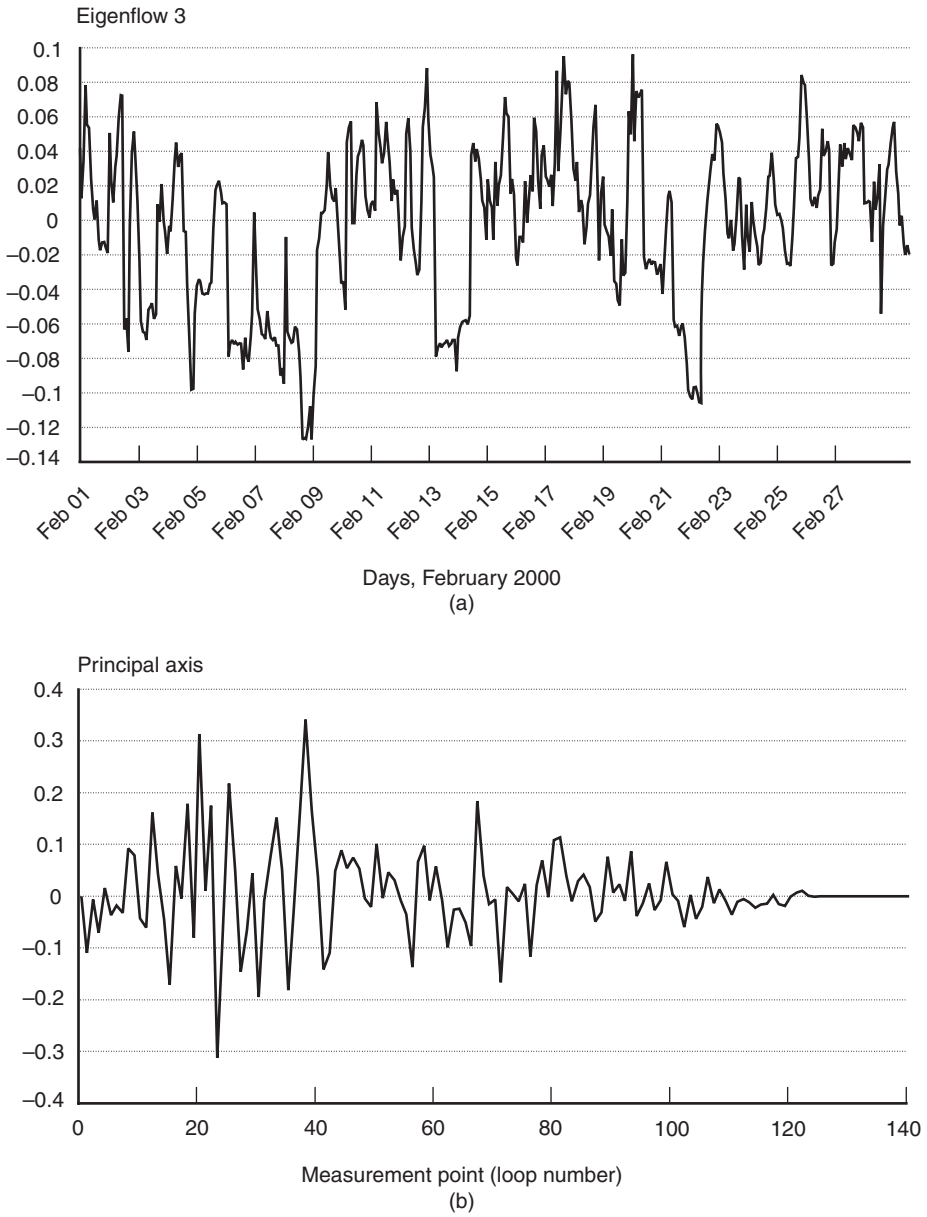


FIGURE 2 Graphical Representation of (a) an Eigenflow and (b) its Corresponding Principal Axis



temporal trends are those corresponding to traffic flow at number 20 and traffic flow at number 23 (see figure 1).

Based on the definition of eigenflows and principal components (axis) in the second section, the negative sign of many entries in some eigenflows, such as in eigenflow 3, denotes that the corresponding common temporal variation pattern is negatively correlated with some of the measured traffic flow time series. Respectively, the negative sign of many entries in the corresponding principal components indicates the negative value of the covariance of the traffic flow measured at a specific traffic flow measurement point, such as the measurement point number 23 (see figure 2(b)), with the other traffic flows. Namely, an increase of the traffic flow rate at measurement point number 23 would result in the reduction of the traffic flow rate in the other measurement points. This kind of analysis helps identify different locations in the network as well as periods of the day, days, or weeks wherein a particular traffic flow has a large impact on the aggregate network traffic conditions.

Measuring the Variability of Individual Flows

Based on the definition given in the first section, each eigenflow can be considered as a building block of the overall dynamics pertaining to each traffic flow. Thus, the variability of individual flows can be determined with regard to the number of significant eigenflows that constitute them. The number of significant eigenflows refers to the number of entries in the corresponding rows of the principal matrix V that are significantly different from zero. There exists a threshold that is equal to $1/\sqrt{m}$ when a row of V has all entries equal, which implies a perfectly equal mixture of all eigenflows, taking into account that the columns of V have unit norm (Lakhina et al. 2004). Then, the number of significant eigenflows is obtained by counting how many entries in each row of matrix V exceed this threshold in absolute value. This approach allows determining the least required number of significant eigenflows, dependent on the sample size, which can provide a plausible reconstruction of each traffic flow (see below), based on equation (4).

Figure 3(a) illustrates the number of significant eigenflows that constitute traffic flows, as this is expressed by the Cumulative Density Function (CDF) of the number of entries per row of V that exceed the above threshold. The curve indicates that no traffic flow is composed of more than 45 significant eigenflows. In particular, it can be observed that 50% of traffic flows are composed of less than 30 significant eigenflows, and more than 30% of traffic flows are composed of less than 20 significant eigenflows. In addition, figure 3(b) presents the histogram of significant eigenflows that constitute traffic flows. This histogram shows that the class interval containing up to 5 significant eigenflows appears most frequently among traffic flows, with the class intervals containing 6 to 10 and 11 to 15 significant eigenflows to follow in order. These results clearly demonstrate that the temporal evolution of most traffic flows can be explained by only a small number of common underlying sources of variability.

Figure 4 shows the number of significant eigenflows with respect to the monthly average daily (for the respective period) traffic flow rate measured over the different detector locations. By and large, the results demonstrate that there is a relationship between the size of a traffic flow and the eigenflows that comprise it. More specifically, the larger traffic flows tend to be composed primarily of a large number (>20) of significant eigenflows (see cluster at the right-hand side of figure 4 separated by a dashed line), in comparison to the smaller traffic flows, which are basically composed of a small number (<20) of significant eigenflows (see cluster at the left-hand side of figure 4). Consequently, the temporal variation of the larger flows has the most significant contribution to the long-term dynamics of the aggregate network traffic in relation to the variation of the other flows.

Measuring the Variability of Aggregate Network Traffic

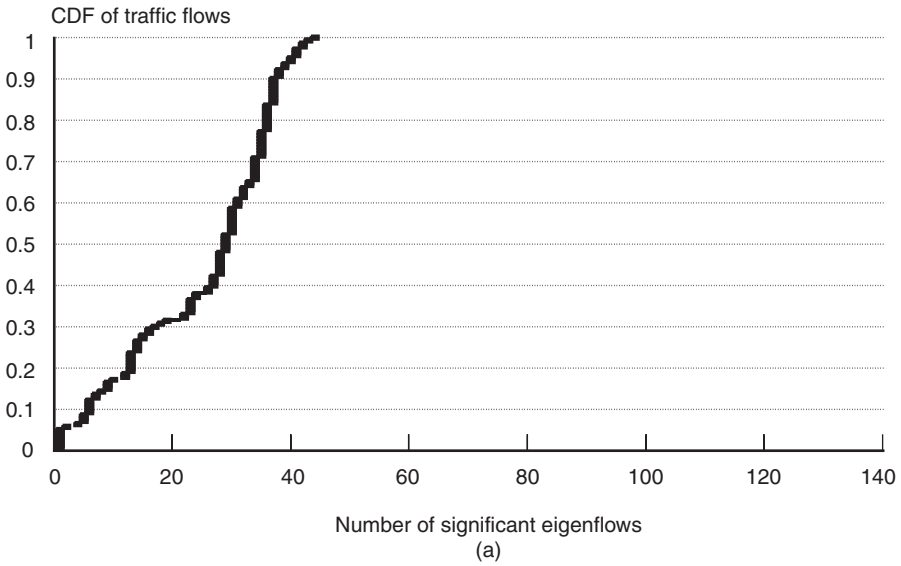
As explained in the second section, the singular values denote the overall variation attributable to each particular principal component. Hence, the order of magnitude of singular values can provide a plausible measure of the extent of variability in aggregate network traffic. Figure 5(a) shows the plot of singular values, in order of decreasing magnitude, corresponding to each traffic flow. This plot clearly demonstrates that the variability in traffic flow can be attributed to only a very small number of eigenflows, that is, common patterns of temporal variation. More specifically, the vast majority of traffic variability is contributed by the first few eigenflows, as signified by the sharp knee of the curve between the third and the eighth singular value. This result provides evidence of the small variability (spread) of the aggregate network traffic in the long run.

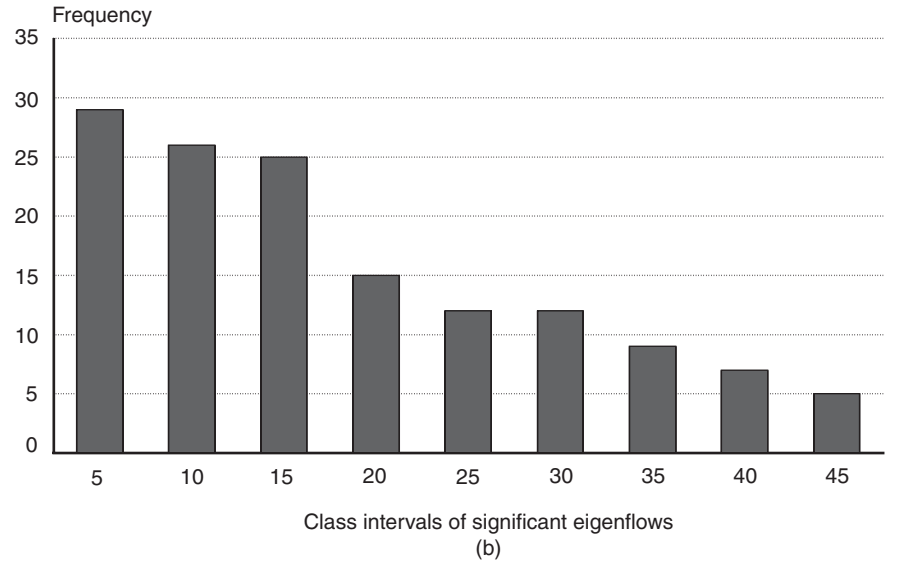
Given the effect of the size of traffic flow on the variability of individual traffic flows, as described in the previous subsection, the effect of the mean traffic flow rate on the small variability of the aggregate network traffic is also investigated here. For this purpose, a zero-mean normalization is applied, which denotes that all measurements of each time series X_i are subtracted from the corresponding sample mean so that their average is zero, as follows:

$$X_i' = X_i - \mu_i, \qquad i = 1,...,m \tag{5}$$

where $\mu_i = \mu(X_i)$ is the sample mean of time series X_i. Figure 5(b) shows the plot of singular

FIGURE 3 (a) Number of Signficant Eigenflows, in Terms of the Cumulative Density Function (CDF) of the Number of Entries in Each Row of the Principal Matrix That Exceed the Threshold and (b) Histogram of Significant Eigenflows

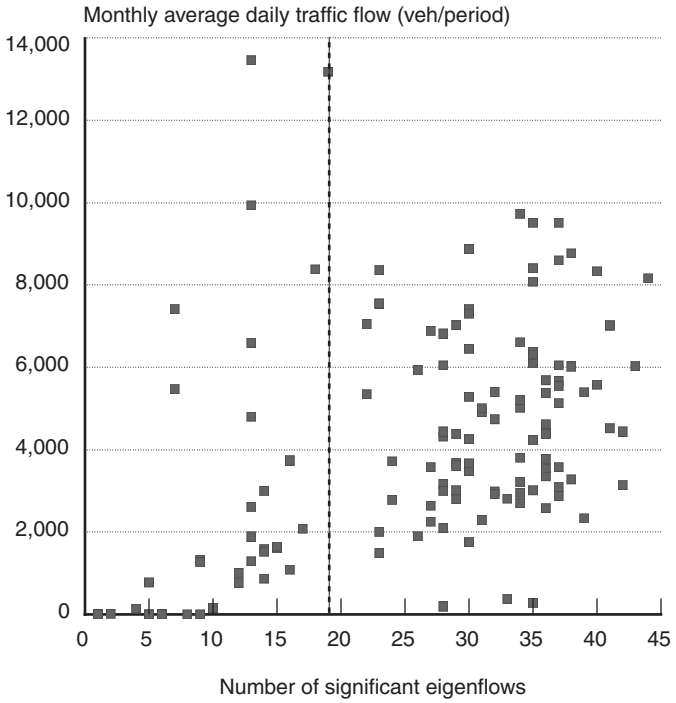




values corresponding to each traffic flow, based on the normalization of traffic flows, as indicated above. In contrast to the case of using the original traffic flows, in the case of using the normalized traffic flows the bulk of variability is signified by a less sharp knee of the curve between the 7th and the 20th singular value. Namely, the relative significance of the first few eigenflows has diminished. This effect can be explained by the fact that a large diversity in the magnitude of the mean traffic flow rate can render the variation of those traffic flows with increased size dominant in comparison to the variation of the other flows. In turn, this leads to a larger diversity between the first few singular values and the remaining singular values.

On the other hand, the fact that the profound majority of variability in traffic flow can still be attributed to only a very small number of eigenflows indicates the dominant role of the remaining effect, which is the effect of the correlation between temporal variation patterns in comparison to the effect of differences in flow size. Therefore, the process of normalization can ensure that the representation of

FIGURE 4 The Number of Significant Eigenflows with Respect to the Monthly Average Daily Traffic Flow Rate



these correlations by eigenflows is not skewed due to differences in the mean traffic flow rate.

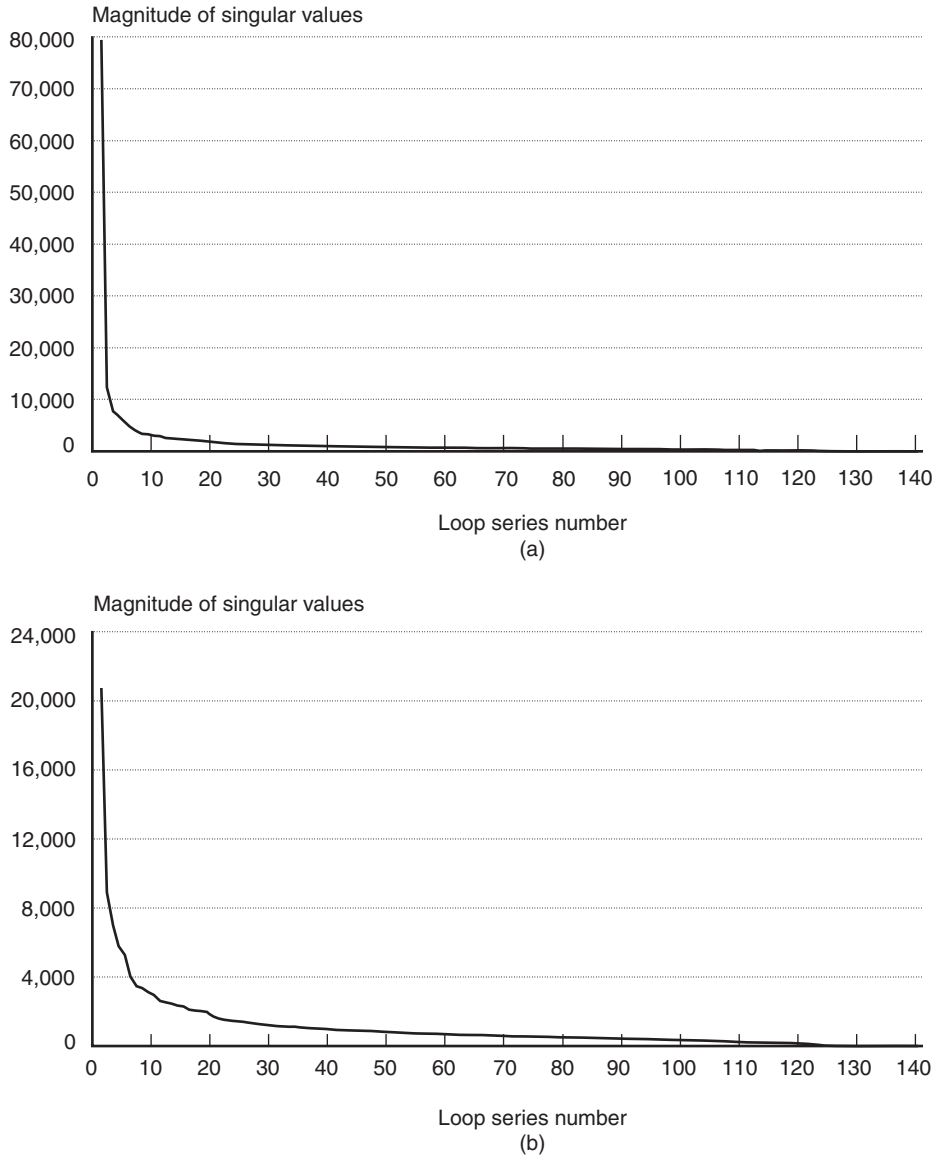
Implication of Variability Measurement for Traffic Data Reconstruction

The fact that only a few singular values can depict the largest portion of the overall variation in aggregate network traffic demonstrates the potential to reconstruct traffic flows or approximate each column of the measurement matrix X, using a considerably smaller number of dimensions. The traffic flows reconstructed by using the whole set of significant eigenflows, on the basis of equation (4), are found to approximate the original (normalized) traffic flows without statistically significant differences at least at the 95% confidence level of the Student t-test statistic. This outcome indicates the correctness of the previously described method for selecting threshold values to determine the number of significant eigenflows composing each traffic flow. Moreover, traffic flow at measurement point number 3 (see figure 1) is randomly selected here to be approximated by using a number of q=5 dimensions (see figure 6(a)). This traffic flow is composed of 30 significant eigenflows. The graphical analysis shows that the temporal pattern of the reconstructed traffic flow remarkably resembles the temporal pattern of the original traffic flow.

Figure 6(b) shows the statistical analysis of the regression of the reconstructed traffic flow to the original traffic flow. The reconstructed flow generally underestimates the original flow, as it denotes the value of the slope of the linear trend line, which is lower than unity (<1.0). This underestimation refers mainly to traffic flows of lower size (<400 veh/15-min), as this is implied by the outliers corresponding to such flow sizes. This outcome indicates that the first 5 (most significant) eigenflows, which are employed in the reconstruction process, can better capture the temporal variation of larger traffic flows in comparison to the remaining eigenflows.

On the other hand, the R^2 value, which represents the squared multiple correlation between the two datasets, indicates that the reconstructed flow data can capture approximately 80% of the systematic variation contained in the original flow data. The above results emphasize the ability of the proposed method to concentrate on a very small set of common sources of temporal variability in order to describe the complexity of traffic flow. In turn, this

FIGURE 5 Plot of Singular Values for (a) Traffic Flows and (b) Normalized Traffic Flows



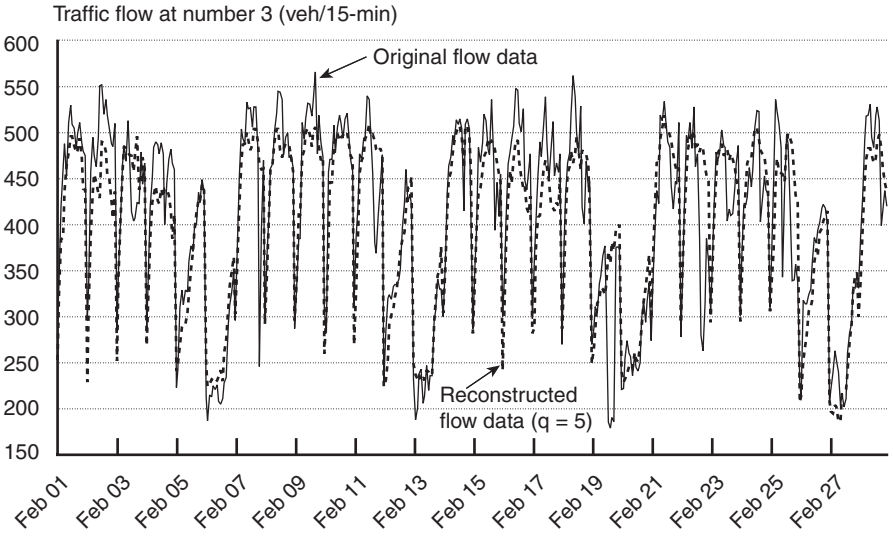
facilitates the deeper understanding and a more plausible interpretation of the factors contributing to the long-term evolution of the main characteristics of urban network traffic.

Implication of Variability Measurement for Traffic Flow Prediction

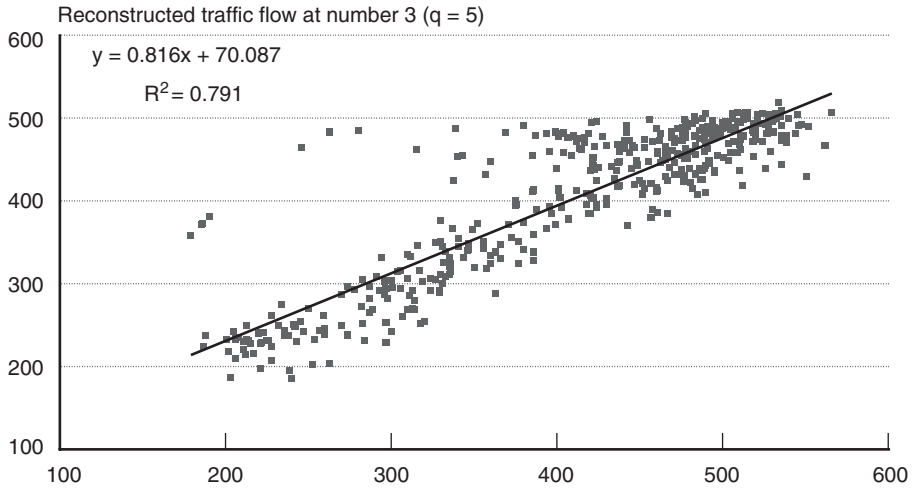
The outcome of the previous subsection (that only a very small set of eigenflows is sufficient for the plausible reconstruction of a traffic flow) emphasizes the need for investigating the potential of the PCA method to approximate future traffic flows. This task is addressed by analyzing data that were not part of the input to the PCA procedure. More specifically, the PCA method is applied to the traffic data, denoted as X^{w1}, measured over a time period spanning the week Monday, Feb. 7, 2000, through Sunday, Feb. 13, 2000, to obtain the principal components $\{v_i\}_{i=1}^{m}$. Subsequently, these principal components are also used to approximate (predict) traffic flow data, denoted as X^{w2}, over the next week spanning Monday, Feb. 14, 2000, through Sunday, Feb. 20, 2000.

The error of approximating a typical traffic flow, that is, traffic flow at number 3, as used in the previous subsection, is investigated for both the first and second week, based on the principal components obtained from the PCA of the first-week data. In

FIGURE 6 (a) Reconstruction of a Typical Traffic Flow Using 5 Principal Components and (b) Statistical Analysis of the Reconstructed Traffic Flow



(a)



(b)

order to investigate the degree to which the flows of the second week preserve the hierarchical structure of temporal variability pattern pertaining to the flows of the first week, the approximation is carried out by using different dimensions (amounts of principal components), that is, q=5, q=10, q=20 and q=40. The traffic flow at number 3 corresponding to the first week, denoted as X_3^{w1}, is composed of 33 significant eigenflows, while the traffic flow at number 3 corresponding to the second week, denoted as X_3^{w2}, is composed of 32 significant eigenflows. The approximation error is measured through (a) the Mean Relative Error (MRE) of the approximation corresponding to each week, which is given as $\left|X_3^{w1} - \tilde{X}_3^{w1}\right| / X_3^{w1}$ and $\left|X_3^{w2} - \tilde{X}_3^{w2}\right| / X_3^{w2}$ respectively, where $\tilde{X}_3^{w1}$ and $\tilde{X}_3^{w2}$ are the reconstructed flows for the first and the second week, and (b) the value of the R^2 coefficient.

Table 1 shows the approximation error, as expressed by the measures of MRE (%) and R^2, for the reconstructed flows of the first and the second week, based on the principal components obtained

TABLE 1 Values of MRE Resulting from the Reconstruction of a Typical Traffic Flow (in percent)
(Values are across 2 successive weeks using principal components of the first week)

Number of dimensions q	5	10	20	40
First-week data	[1]6.40	[2]5.64	35.04	3.25
	(0.935)	(0.952)	(0.965)	(0.988)
Second-week data	[1]7.39	[2]6.50	[3]6.38	6.26
	(0.904)	(0.926)	(0.934)	(0.941)

[1, 2, 3] Pairs with no statistically significant differences at the 95% confidence level of the t-test statistic

Note: Values in parentheses indicate R^2.

from the PCA of the first-week data. The results demonstrate that, when using the same number of dimensions (q=5), the application of the PCA on a shorter term dataset, such as that of one week, results in a more accurate approximation of the original flows (R^2=0.935) in comparison to the application on a longer term dataset, such as that of four weeks (R^2=0.791). This outcome can be attributed to the fact that longer term data typically involve more sources (larger spread) of temporal variability in the network.

The approximation of $\tilde{X}_3^{w2}$ based on the first-week principal components, as well as the approximation of $\tilde{X}_3^{w1}$, resulted in low MRE values (<10.0%) and high R^2 values (>0.90). The differences between the MRE values that resulted from the two approximations were not found to be statistically significant at the 95% confidence level of the t-test statistic for all sets of principal components used, except for the case using 40 principal components (see footnote of table 1). Hence, the first-week principal components can be well used to approximate traffic flows of the second week, such as $\tilde{X}_3^{w2}$. Moreover, the loss of the predicting power of the first-week principal components can be attributed to the last few eigenflows, which are mostly related to smaller size and higher variability traffic flows. The results generally provide evidence of the increased temporal stability of the hierarchical pattern of traffic variations from one week to the next. Thus, they indicate the potential of using the first few eigenflows of the previous week to consistently reproduce, with a reasonable accuracy, most systematic features of the traffic flow of the next week.

DECOMPOSITION OF EIGENFLOWS AND APPLICATIONS

Method for Decomposing Eigenflows

Each eigenflow can be decomposed into nonstationary and stationary components, according to the nature of variability in traffic flow. For this purpose, each eigenflow is modeled here as an unobserved-components time series in state space form (Koopman et al. 1999), so that they enable the smoothing estimation of both nonstationary and stationary components. The nonstationary component refers to nonstationary variation (or changes) of the eigenflow mean, and it reflects periodicities, namely periodic trends, in traffic flow. These time-varying trends are due to diurnal cycles in travel demand, differences in traffic conditions among weekdays, as well as between weekdays and weekends. This component is calculated here with the state smoothing of each eigenflow, which captures changes in the level or trend of traffic variability in the long run.

The stationary components refer to structural breaks and outliers. These are typically expressed with isolated values, which are located outside a band of, for example, ±2 standard errors (SE) from the trend line, that is, the smoothed eigenflow mean. The structural breaks reflect occasional bursts and dips in the level of traffic variability. These breaks correspond to stochastic and transient changes, such as traffic phase transitions, pertaining to the physical dynamics of (recurrent) congestion conditions. The outliers reflect noise, that is the remaining random variation in traffic data. They can be attributed to extreme or unusual traffic-related events, principally related to nonrecurrent congestion dynamics, such

as demonstrations, road works, sportive events, emergency situations, accidents, or other incidents.

The existence, in terms of their statistical significance, of each of these two types of stationary variation is identified here by applying the *t*-test statistic to the results obtained from the disturbance smoothing of each eigenflow. The solution of both the state smoothing model and the disturbance smoothing model is carried out by using an appropriate maximization routine written in the *Ox* matrix programming language (Doornik 2002). Further information on the analysis of time series using state and disturbance smoothing models can be found in Durbin and Koopman (2001).

Practical Demonstration of the Method

For demonstration purposes, the proposed method for the decomposition of eigenflows is implemented here for the same typical eigenflow used in the previous section, that is, eigenflow 3, at a finer level of temporal resolution, that is, a period of one week spanning from Monday, Feb. 14, 2000, to Sunday, Feb. 20, 2000. Based on the results of the previous section, the normalized traffic flows are used for the estimation of the temporal trend of the eigenflow. The usage of these data prevents the effect of possible bias caused by differences in flow size between various time intervals of the day as well as periods of successive days-of-the-week. Figure 7a presents the estimation of the temporal trend of the given eigenflow together with the corresponding band of ±2SE from the estimated trend line. In addition, figure 7b illustrates the distinction of structural breaks and outliers, which correspond to the given eigenflow, and their statistical significance, as determined by the range defined between the upper and the lower confidence level, through the *t*-test statistical analysis of these two types of stationary variation.

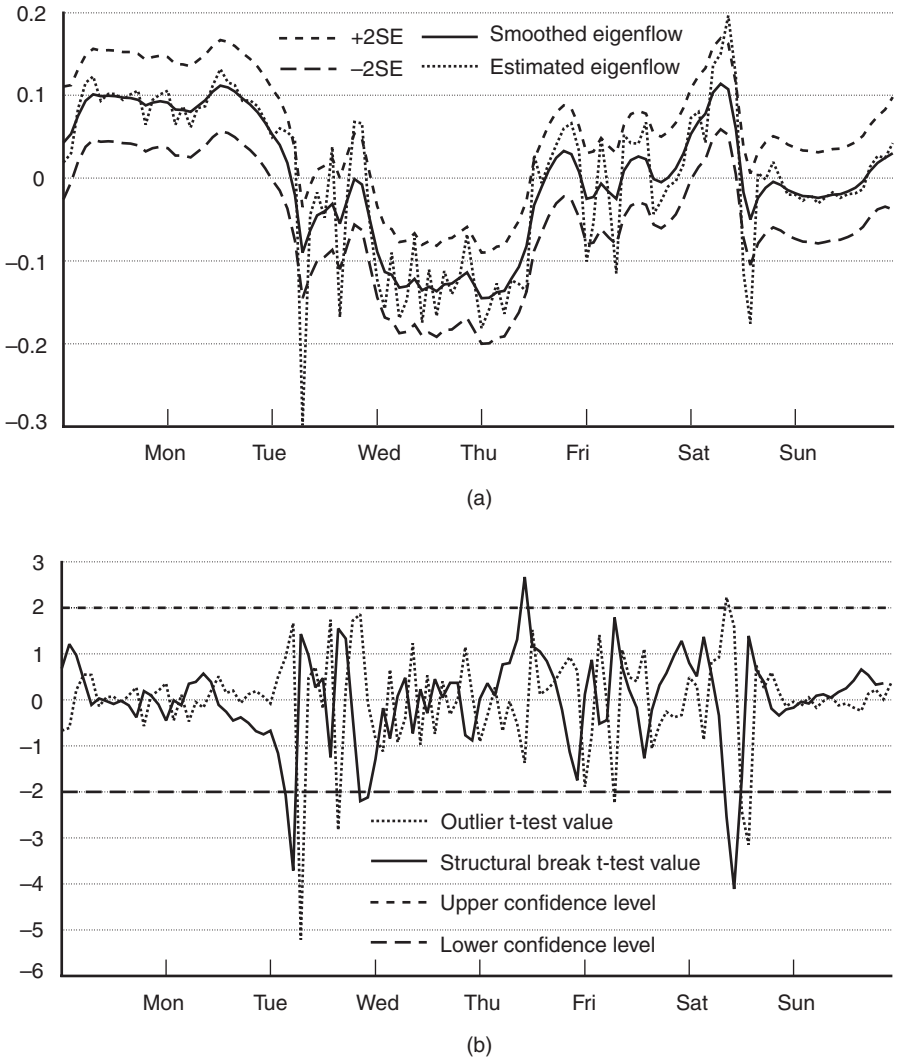
The process which was adapted here, referred to as temporal trend thresholding, through the suitable selection of a band with magnitude ±2SE, appears to provide a reasonable distinction between stationary variations and nonstationary changes from the eigenflow mean. This is clearly demonstrated by the fact that the selection of such a threshold or bandwidth can capture the existence of breaks and outliers because the values that are kept out of this band (see figure 7a) correspond to statistically significant stationary variations (figure 7b), based on the *t*-test statistic. In the present case, these short-lived, statistically significant outliers are mainly due to bus breakdowns and local accident situations. The above process can also provide information on the extent to which statistically significant breaks and outliers affect changes in the long-term trend of a common variation pattern.

This type of analysis is particularly helpful for understanding the characteristics of urban traffic, provided that most traffic flows can be sufficiently represented through only a small number of eigenflows, as shown in the previous section. On the one hand, the temporal trend thresholding is suitable for the long-term analysis of the expected or predictable variations of traffic, that is, those variations restricted within the band of ±2SE. The determination of the periodic trends of each eigenflow makes it possible to identify the extent to which the variability in traffic flow is predictable within time periods, such as those from week to week.

Furthermore, these results have implications for the verification or validation of traffic assignment and simulation models used to provide traffic predictions. The temporal trend thresholding enables the definition of bounds in which a long-term prediction can be considered as statistically reliable, or, otherwise, an extreme or erroneous forecast, which should be ignored or set equal to the upper or lower statistical bound of the corresponding point in time. In addition, this procedure can help identify whether traffic predictions in some links are more error-prone than others.

On the other hand, the process described here proposes a new scheme for identifying the location and the time of occurrence of statistically significant stationary variations in traffic flow, related to unusual or extreme events, as described in the previous subsection. The operation of the proposed scheme can be updated periodically (e.g., from week to week) in an automated manner by simultaneously reprocessing traffic flows measured over different locations in the network while it enables the statistical analysis of different types of stationary variation, such as structural breaks and outliers. For these reasons, this scheme can be considered as a more rigor-

FIGURE 7 Decomposition of a Typical Eigenflow by Use of (a) Temporal Trend Thresholding and (b) *t*-test Statistical Analysis of Structural Breaks and Outliers



ous and practically useful approach than the method of detecting outliers through simply comparing individual flows to some average traffic pattern obtained from past measurements over a given location and period of time (e.g., a week). Moreover, this process introduces a methodology for deriving a set of different models for local traffic prediction across multiple timescales by taking into account the fact that the traffic in different parts or links of the network may experience different rates and types of variation as time progresses.

CONCLUSIONS

This paper describes the analysis and interpretation of the variability in urban traffic flow by processing one-month traffic detector data corresponding to a realistic large-scale arterial network. The method of principal component analysis was found to provide a plausible and powerful tool for the purposes of the present study. Specifically, the PCA enables the identification of eigenflows, which denote common patterns of temporal variability, according to their contribution to the aggregate network traffic. Despite the underlying complexity in the phenomenology of urban traffic structure, the findings suggest that the spatio-temporal variation in traffic flow in such a network can be represented by only a small set of eigenflows. This small variability can be attributed to the increased correlation between temporal variation patterns and the presence of periodic

trends in these patterns, and it was found to carry useful implications for the updated prediction of traffic flow patterns, for example, from one week to the next.

Moreover, the statistical analysis of the calculated eigenflows allows for the presence of stationary variations, namely, breaks and outliers. The identification of such variations is particularly valuable for supporting real-time network operations, including detection of traffic anomalies and incidents. In addition, the proposed methodology offers a valuable tool to manage stored aggregate traffic flow data in large-scale urban networks for planning purposes. Such purposes can encompass the assessment of traffic responsive control strategies, the verification of traffic assignment and simulation models used to represent the variation in traffic patterns, the evaluation of the traffic network performance, and its impact on the environment.

REFERENCES

Davis, G.A. 1997. Accuracy of Estimates of Mean Daily Traffic: A Review. *Transportation Research Record* 1593:12–16.

Doornik, J.A. 2002. *Object-Oriented Matrix Programming Using Ox,* 3rd ed. London, England: Timberlake Consultants Press. Available at: http://www.doornik.com.

Durbin, J. and S.J. Koopman. 2001. *Time Series Analysis by State Space Methods*. Oxford, England: Oxford University Press.

Koopman, S.J., N. Shephard, and J.A. Doornik. 1999. Statistical Algorithms for Models in State Space Using SsfPack 2.2. *Econometrics Journal* 2(1):113–166.

Lakhina, A., K. Papagiannaki, M. Crovella, C. Diot, E.D. Kolaczyk, and N. Taft. 2004. Structural Analysis of Network Traffic Flows. *Proceedings of the International Conference on Measurements and Modeling of Computer Systems SIGMETRICS.* Edited by E.G. Coffman Jr., Z. Liu, and A. Merchant. New York, NY: ACM.

Meyer, C.D. 2000. *Matrix Analysis and Applied Linear Algebra.* Philadelphia, PA: SIAM.

Rakha, H. and M. Van Aerde. 1995. Statistical Analysis of Day-to-Day Variations in Real-Time Traffic Flow Data. *Transportation Research Record* 1510:26–34.

Sharma, S.C., B.M. Gulati, and S.N. Rizak. 1996. Statewide Traffic Volume Studies and Precision of AADT Estimates. *Journal of Transportation Engineering* 122(6):430–439.

Stathopoulos, A. and M.G. Karlaftis. 2001. Temporal and Spatial Variations of Real-Time Traffic Data in Urban Areas. *Transportation Research Record* 1768:135–140.

Treiber, M. and D. Helbing. 2002. Reconstructing the Spatio-Temporal Traffic Dynamics from Stationary Detector Data. *Cooperative Transportation Dynamics* 1:3.1–3.24.

Weijermars, W.A.M. and E.C. van Berkum. 2004. Daily Flow Profiles of Urban Traffic. *Urban Transport X*. Edited by C.A. Brebbia and L.C. Wadhwa. Southampton, UK: WIT Press.