Revealing a Hidden, Stable Spectral Structure of Urban Vehicular Traffic

Fan Bai
General Motors Global R&D

Bhaskar Krishnamachari
University of Southern California

2021 IEEE Vehicular Networking Conference (VNC) | 978-1-6654-4450-7/21/$31.00 ©2021 IEEE | DOI: 10.1109/VNC52810.2021.9644654

***Abstract*—A deeper understanding of urban vehicular traffic is important to enable better design and evaluation of future vehicular and cellular communication networks. In this paper, we study the presence of spectral structure in urban vehicular traffic. By analyzing publicly available sets of fleet vehicle mobility traces obtained from two real-world deployments that consist of more than 2,000 taxis in Shanghai and Beijing respectively, we reveal the existence of a stable, low-dimensional spectral structure in vehicular networks, which was often unnoticeable when using classic spatio-temporal data analysis. This stable spectral structure not only significantly simplifies the representation of high dimensional transportation data, but also offers interpretable insights into urban mobility patterns. Leveraging the stability of spectral structure, we demonstrate that the spectral structure analysis could effectively tackle practical problems in the field of transportation research, such as traffic anomaly detection.**

I. Introduction

To improve traffic management and operations in metropolitan areas, Intelligent Transportation Systems (ITS) rely on sufficient and high-quality field data – vehicle density, vehicle flow speed, and occupancy, among others – to provide proper traffic controls. The ITS field data, collected either by roadside transportation sensors[2] (e.g., induction loops, traffic camera, or Radar/LIDAR sensors) or through floating car sensors[3], [4] (e.g., GPS, RFID, and transponders), describes transportation phenomena across both space and time. We call such data *spatio-temporal data*.

The vast majority of research works on ITS data analytic were conducted on the original spatio-temporal domain [1]. For example, researchers leverage the established spatio-temporal relationship of traffic flows [6] to measure traffic pattern variation over time, determine Origin-Destination (OD) matrix, identify the functionality of different geographic regions, and figure out the influx and outflux patterns within a city [1]. However, this mainstream spatio-temporal domain approach faces a number of significant challenges that hinder it from achieving its goals, especially when being applied to a large volume of high dimensional ITS data which is often noisy and garbled [7]. These technical challenges call for a change in ITS data analytic research, motivating us to find a new approach that could systematically study high dimensional ITS data collected from future smart cities. We are motivated to study ITS data not only from the perspective of the design and control of intelligent traffic management systems and future autonomous vehicles, but also from the perspective of helping to improve the design and evaluation of future communication networks including vehicular networks and cellular networks. For example, getting a deeper understanding of ITS data on vehicular traffic from real urban environments can make it easier to simulate network traffic to and from connected vehicles at a city-scale. And in turn, such simulations could help in the design of emerging protocols and standards related to V2X, 5G, and beyond [28].

Somewhat to our surprise, we found that there are very few research works that attempt to study transportation traffic from a spectral analysis perspective. To bridge this gap, in this paper, we propose to examine high dimensional ITS data from a new perspective, and focus on analyzing the spectral structure of urban vehicular traffic. While doing so, we aim to develop a generic framework to systematically study the spectral characteristics of ITS data, which are often either unnoticeable or ignored in spatio-temporal domain. We attempt to answer the following questions:

1) Does urban ITS data show the existence of a stable, low-dimensional spectral structure?
2) If yes, what are the properties of this spectral structure?
3) If yes, how do we utilize this spectral structure for solving practical problems?

Using Principle Component Analysis (PCA), a spectral analysis tool, we analyze two publicly available sets of taxi fleet mobility traces, with each trace consisting of more than 2,000 taxis. Through this analysis, we indeed discover the existence of such a spectral structure in the cases of Beijing and Shanghai taxi traffic. Interestingly enough, we also find that the spectral elements are intepretable, revealing a deep insight of movement patterns. The spectral structure of vehicular traffic is stable over time and over space, forming a solid foundation for data reconstruction. We demonstrate that the spectral characteristics are not only capable of annotating the hidden structure of urban mobility patterns, but also are able to identify traffic anomaly events.

The contributions of this paper are summarized as two folds: First, by using large-scale empirical measurement traces, our study is the first to reveal a stable, low-dimensional spectral structure that governs the fundamental laws of vehicular traffic patterns. We attribute this spectral structure to human's structured habits in traveling. Second, by leveraging this stable spectral structure of urban taxi traffic, our study successfully isolates traffic anomalies from regular traffic events. This is achieved by separating Gaussian noise and spike principle components, in spectral domain, from their periodic counterparts.

978-1-6654-4450-7/21/$31.00 ©2021 IEEE

II. Background on Spectral Analysis

Spectral analysis is used for reducing the dimensionality of high dimensional data which has inherent redundancy. As an example, Principle Component Analysis (PCA), a prominent spectral analysis method, transforms data samples from their original coordinate system into a new coordinate system, in which each of its orthogonal dimensions successively maximizes the statistical variance of the original data.

Spectral analysis not only achieves the goal of dimension reduction, but can also increase the data interpretability by revealing the hidden low-dimensional structure, which is often ignored in its original coordinate system. Further, it can be used as a tool for better interpolation of missing data.

A. PCA Decomposition

To filter out the noise and reveal hidden dynamics, through its coordinate transformation process, PCA finds the most meaningful coordinate system to re-express a noisy, garbled data set. Each dimension of the new coordinate system (it is called *principle component*) points in the direction of maximum variation remaining in the data set, given the variance already accounted in the previously identified principle components. For instance, the first principle component captures the dimension that reflects total variance of the original data on a single dimension of this new coordinate system; the next principle components thus characterize the maximum residual variance for the remaining orthogonal dimensions of the new coordinate system, respectively. To put it into mathematical terms, by solving an eigenvalue and eigenvector problem of the covariance matrix of measured data samples, we can find such new principal components.

Principle Component (PC) Vectors. Let X be the vehicle traffic measurement matrix in the original spatio-temporal space; X is a two-dimension matrix with t time slots and p geographic regions. We consider each row vector of traffic matrix $X \in \Re^{t \times p}$ as a data point in $\Re^{p}$, and thus X_p contains t data points.

In the transformed new coordinate system, we represent principle component vectors as $\{v_i\}_{i=1}^{p}$. For the first principle component vector v_1, it captures the maximum variance of the original traffic matrix X in the spatio-temporal space as follows:

$$v_i = \arg \max_{||v||=1} ||Xv||. \tag{1}$$

For remaining principle components v_i ($i \geq 2$), it captures the maximum residual variance of the traffic matrix, which excludes the variance accumulated by the first $i-1$ principle components, as follows:

$$v_i = \arg \max_{||v||=1} ||(X - \sum_{k=1}^{i-1} X v_i v_i^T) v||. \tag{2}$$

Eigenvalues of PCA. As learned from matrix theory, principle component vectors $\{v_i\}_{i=1}^{p}$ are in fact the eigenvectors of the matrix $X^T X$. Therefore, the eigenvalues $\{\lambda_i\}_{i=1}^{p}$ could be obtained by solving the following formulation:

$$X^T X v_i = \lambda_i v_i \tag{3}$$

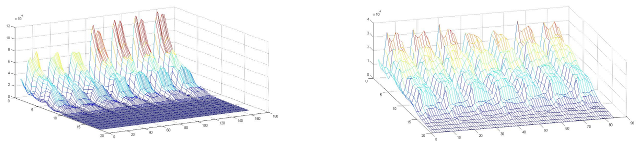
where non-negative $\{\lambda_i\}_{i=1}^{p}$ follows a descending order, such as $\lambda_1 \geq \lambda_2 ... \geq \lambda_p \geq 0$.

We apply PCA decomposition towards both temporal dimension $\Re^{t}$ and spatial dimension $\Re^{p}$ of traffic measurement matrix $X \in \Re^{t \times p}$, revealing its temporal and spatial characteristics, respectively, in Section IV.

III. Traffic Measurement Matrix

Our analysis is focused on taxi density matrix, based on empirical GPS traces collected in Beijing and Shanghai. Though only two specific data sets are selected, it should be understood that our spectral analysis methodology is not limited to particular transportation methods or particular cities; our study develops a generic framework to investigate the fundamental principles that govern the urban traffic patterns.

Data Collection. The first taxi data set in our analysis was collected from Beijing, China, and consists of 2,721 taxis over two weeks, May 1 - May 14, 2009. The secondary set was collected from Shanghai, China on January 31, 2007 - February 27, 2007 (1 month), and is composed of over 2,439 taxis. The logged data in the data sets includes vehicle ID, time-stamp, longitude and latitude coordinates, speed and heading, and occupancy status of a taxi vehicle. Due to cellular cost constraint, each taxi only afforded to report its mobility trajectory every 15 seconds (with passengers on board) or every 60 seconds (without passengers). We use the geographic boundary of each district of Beijing (or Shanghai) to classify a taxi's current location to its appropriate district.



(a) Beijing

(b) Shanghai

Fig. 1: The Taxi Density Matrix of Beijing and Shanghai in a Week. The x-axis represents the time slot (in a unit of one hour), the y-axis represents the districts (in a decreasing rank of their taxi populations), and the z-axis shows the taxi density $X(t, p)$ at time slot t in district p.

Traffic Measurement Matrix. We use taxi density matrix as an example to illustrate the spectral analysis method. The taxi density matrix $X \in \Re^{t \times p}$ is a 2-dimensional matrix defined as the density of taxis in a particular region p at time slot t. This matrix reflects the urban traffic conditions (e.g., whether region p is a busy district at time t?). Fig. 1 shows such examples in Beijing and Shanghai, based on data collected in a week. In the case of Beijing, we intentionally choose a week with three consecutive weekend days (including the International Labor Day as a national holiday).

The vehicle density in Beijing (Fig. 1(a)) fluctuates in a daily base, ranging from sparse traffic at night to high volume in the rush hour. We observe that taxis are unevenly distributed over space: Central Business Districts (CBDs) have a much higher concentration of taxis when compared to rural regions, reflecting the geographic functions within a modern metropolitan. Similar observations are made in Shanghai (Fig. 1(b)). In the remaining part of this paper, unless otherwise noted, we omit the results of Shanghai traces for brevity.

IV. Spectral Analysis of Traffic Measurement Matrix

In this section, by following the methodologies laid out in Sec. II, we conduct a PCA analysis of traffic data in Sec. IV-A and reveal its spectral structure in Sec. IV-B.

A. Spectral Analysis via PCA

Initial PCA Results. PCA principle components (PCs) capture the new orthogonal dimensions of transformed coordinate system. Fig. 2(a) and Fig. 2(c) provide an initial clue of three phases in the PCA spectral domain: an initial phase of periodical patterns, an intermediate phase rich with random Gaussian-type noise, and a final phase featuring spike-like noise. We will further examine this initial observation more carefully in Fig. 4.

With PCs being determined, eigenvalues reflect the weights of each principle component. Fig. 2(b) and Fig. 2(d) illustrate the weight of different PCs in Beijing and Shanghai traces, respectively. The initial phase of (periodic) PCs obviously makes a dominating contribution when compared to other PCs.

Low Dimensionality of Urban Traffic. To further understand how these first few PCs play a dominating role in a quantitative fashion, we plot the scree plot for both Beijing and Shanghai taxi traces in Fig. 3. This scree plot summarizes how much residual statistical variance is captured by each orthogonal PCA principle component. As shown in Fig. 3, the first few principle components dominate the statistical variance of urban traffic matrix. In Beijing trace, the first three PCs dominate the 99% variance; on the other hand, it is slightly relaxed in the case of Shanghai taxi trace, in which the first five PCs are the dominating factors that contributed to 99% variance. That is to say, the urban traffic matrix exhibits a hidden structure with very few effective PCA dimensions (in our two cases, less than five), which is significantly lower than that of the original spatio-temporal space.

Principle Component (PC) Categories. As discussed earlier in this section, PCA principle components could be categorized into three major classes: *Periodical PCs*, *Gaussian noise PCs*, and *Spike PCs*. Using a more rigorous method [21], we confirmed that PCA analysis of urban traffic data indeed follows these principle component categorizations. We only present our results obtained from the Beijing trace for brevity.

Periodic PCs. The first few PCs capture the vast majority of statistical variance. Fig. 4(a) and Fig. 4(b) visually show that this category of PCs is periodic. This is not surprising, because the dominant factor of traffic matrix variance is the periodical daily traffic variation.

Gaussian Noise PCs. The second category of PCs represents random, Gaussian-like noise. Such examples are shown in Fig. 4(c) and Fig. 4(d). We found that, though the mere number of Gaussian noise PCs significantly surpasses the other two categories, but its aggregated residual variance could rather be ignored when compared to the first category.

Spike PCs. The final category of PCs features a short-lived, strong spike in its existence. Fig. 4(e) and Fig. 4(f) shows two such examples, representing the elements of occasional traffic bursts or traffic dips. These spike principle components, rare and sporadic, are distributed across the PCA principle component domains.

B. Spectral Structure of Urban Traffic

Motivated by the observation of low dimensionality in urban traffic data, we hypothesize that *few dominant principle components (in spectral PCA domain) consist a hidden structure of urban traffic patterns, despite the fact that this spectral structure is much more succinct than its counterpart in spatio-temporal domain.* To validate our hypothesis, we investigate whether the spatio-temporal traffic matrix could be reconstructed, with reasonable accuracy, by using this spectral-domain structure.

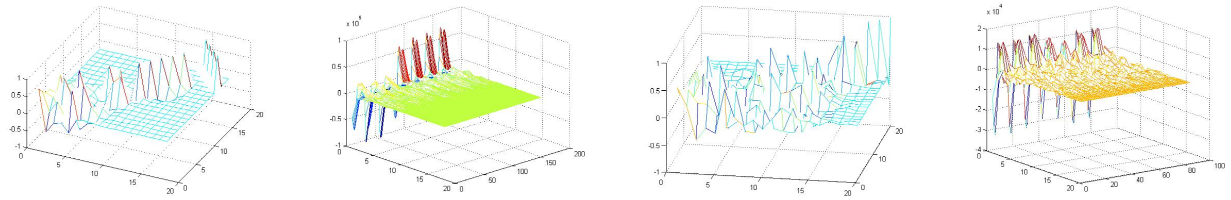
Methodology. The spectral reconstruction is conducted by calculating the best r-rank PCA approximation of the urban traffic matrix X (here, $r \leq p$). Mathematically, it could be expressed as

$$\tilde{X}(r) = \sum_{i=1}^{r} \sigma_i \mu_i v_i^T. \tag{4}$$

This method could be applied for PCA decomposition along either temporal dimension $\Re^t$ or spatial dimension $\Re^p$ of traffic matrix $X \in \Re^{t \times p}$. When determining the r value, we choose periodic PCs only (the first category) due to their dominating roles, while excluding the other two categories (Gaussian noise PCs and spike PCs).

PCA Decomposition along Temporal Dimension. First, we look into the case that PCA decomposition is conducted along temporal dimension. Fig. 5 shows the original version of, as well as the reconstructed version of, vehicle density across different districts in a single week. We choose $r = 3$ in this case. Fig. 5 demonstrates that, in general, the top 3 principle components could accurately characterize the temporal fluctuation of vehicle density. On the other hand, this reconstruction has better reconstruction accuracy in high-population districts than their low-population counterparts (the 1st district vs. the 11th district in Fig. 5); this may be attributed to the fact that a lower variance of density exists in the high-population districts due to more samples (and likewise a higher variance of density in the lower-population districts due to fewer samples.) To better understand the sensitivity of PC dimensions, we plot the reconstruction error as a function of the number of PCs in Fig. 6. As the number of PCs used for reconstruction increases, the reconstruction error reduces but the contributions from extra PCs become marginal after the first three PCs.

PCA Decomposition along Spatial Dimension. We also examine the scenario where PCA decomposition is executed along spatial dimension. Fig. 7 shows the original version of, as well as the reconstructed version of, taxi density in different geographic districts. The reconstruction is obtained by calculating the best r-rank ($r = 5$ in this case) approximation



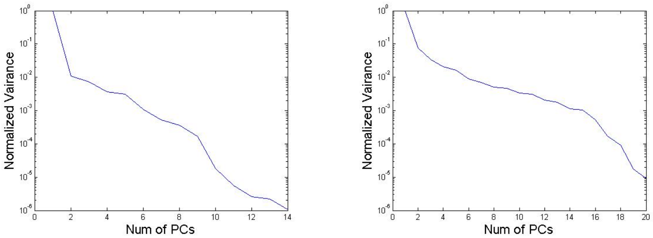
(a) Principle Component (Beijing)

(b) Eigenvalue (Beijing)

(c) Principle Component (Shanghai)

(d) Eigenvalue (Shanghai)

Fig. 2: The PCA Analysis of Taxi Density Matrix (3D Plot for Principle Components and Eigenvalues).



(a) Beijing

(b) Shanghai

Fig. 3: The Scree Plot of Residual Variance vs. Principle Component Dimensions.

of the original matrix X. Again, it is clearly seen that, in the spatial dimension scenario, (only) the top 5 significant spatial PCs could accurately characterize the vehicle density across all districts, except occasional outliers.

By transforming urban traffic data from classic spatio-temporal domain to our advocated spectral domain, we reveal that *there indeed exist a hidden, low dimensional structure of urban traffic patterns*. This revelation not only significantly simplifies the representation of the high dimensional traffic data, but also make it interpretable, as elaborated in the next section.

V. Properties of Spectral Structure

In this section, we further examine the spectral structure of urban traffic and make insightful observations, which are either difficult to obtain, or often ignored, via classic spatio-temporal domain analysis. A rigorous PCA analysis (Sec. II), when properly performed, could reveal interpretable meanings of identified principle components. We show that *spectral structure of urban traffic is not only interpretable, but also stable over time and space.*

A. Temporal Spectral Structure

Interpretation of PCs. We first study the spectral structure along temporal dimension. Fig. 8 illustrates the first 5 PCs, when PCA is applied to temporal dimension. Interestingly enough, we find that the first temporal PC (Fig. 8(a)) resembles the taxi density in the entire city, with a Pearson correlation value of 0.98. It is not out of our expectation, since the first temporal PC is supposed to capture the majority of statistical variance. The second and third PCs seem to represent traffic trend happening in the morning and afternoon (Fig. 8(b)), respectively, with the peak lasting for 4-6 hours. Finally, the fourth and fifth temporal PCs are likely to indicate short-term, rush-hour traffic during morning and evening commutes, with the spike of each lasting for only 2-3 hours.

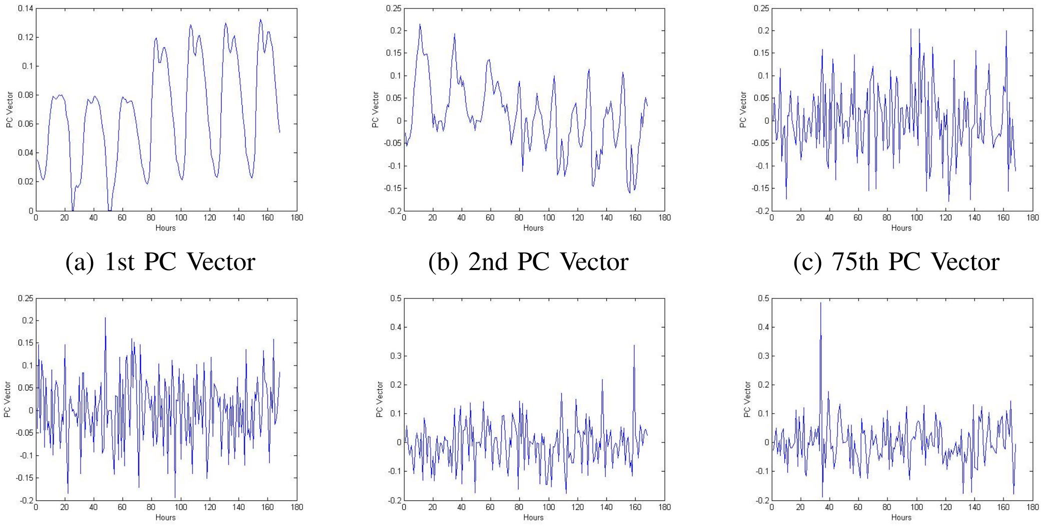
TABLE I: Pearson Correlation of Temporal PCs Between Different Days

PC	Between Weekend	Between Weekday	Between Weekend and Weekday
1st PC	0.9947	0.9989	0.9597
2nd PC	0.8938	0.9619	-0.5855
3rd PC	0.7859	0.8243	0.5839
4th PC	-0.1419	0.7323	-0.0181
5th PC	0.2410	-0.2916	-0.0119

Stability and Repeatability. We observe that the top temporal PCs exhibit strong stability and repeatability, forming a persistent spectral structure governing urban traffic along temporal domain. Table I shows a strong stability of top temporal PCs. The Pearson correlation value between any two weekdays is high till the 4th PC (0.7323), and in the case of any two weekend days, the correlation is strong till the 3rd PC (0.7859). On the contrary, for a given PC, there is either no correlation or negative correlation between a weekday and a weekend day. Interestingly, for the 2nd PC, the Pearson correlation between a weekday and a weekend day is -0.5855, indicating the 2nd PC (reflecting morning traffic) during weekday and weekend is somehow opposite to each other, because people usually get up late during the weekend. (Fig. 9 later shows that is the case.)

To dive into the details, Fig. 9 shows the comparison of the 1st and the 2nd PCs across a week. Again, we find that temporal principle components differ significantly between weekday version (May 4 - May 7) and weekend version (May 1 - May 3). Meanwhile, we observe that, within the weekday category, the 1st and 2nd temporal principle components are almost the same across different days, while the 3rd and 4th temporal principle components still show strong resemblance across different days (we only show the 1st and 2nd principle components in Fig. 9, due to limited space). Such trend exists for other temporal principle components, but their correlation is much weaker, as shown in Fig. 10.

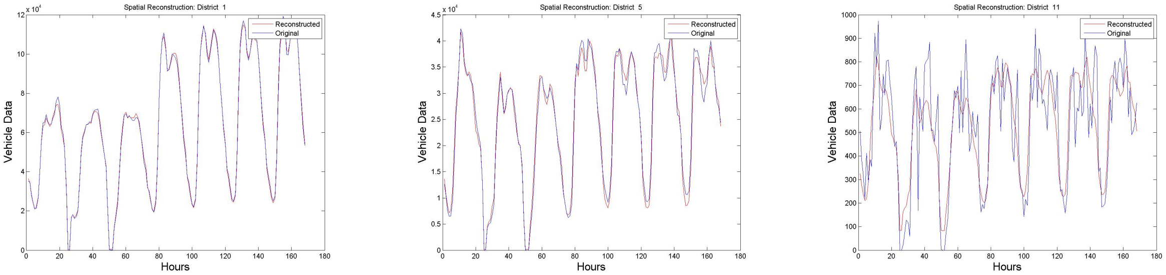
We believe this strong stability is caused by people's daily routine schedules: in the weekdays, people have more structured travel patterns (e.g., a particular original-destination pair at a given time in a repeatable fashion) due to work or school requirements. In contrast, people are more relaxed during the weekend, resulting in a less structured travel pattern.



(a) 1st PC Vector (b) 2nd PC Vector (c) 75th PC Vector

(d) 88th PC Vector (e) 68th PC Vector (f) 99th PC Vector

Fig. 4: Examples of Principle Component (PC) vectors In Beijing Taxi Trace. PCs could be classified into 3 categories: periodic PCs (a)(b), Gaussian noise PCs (c)(d), and Spike noise PCs (e)(f).



(a) 1st District (b) 5th District (c) 11th District

Fig. 5: Reconstruction of Taxi Density in Different Districts with Only 3 Temporal PCs. The x-axis presents the time evolution, and the y-axis indicates the number of taxis in this district at a given hour. It is shown that only top 3 PCs are needed to accurately reconstruct the original data set, thanks to low spectral dimensionality of urban vehicular traffic.

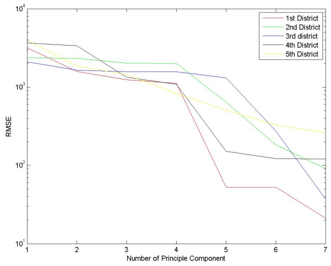
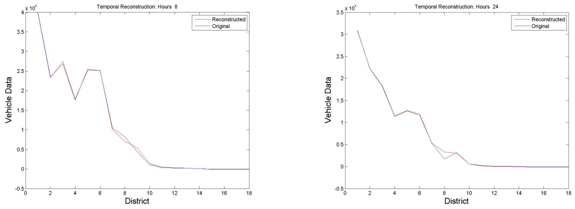


Fig. 6: The Reconstruction Error (Root Mean Square Error, RMSE) vs. Number of Principle Components.

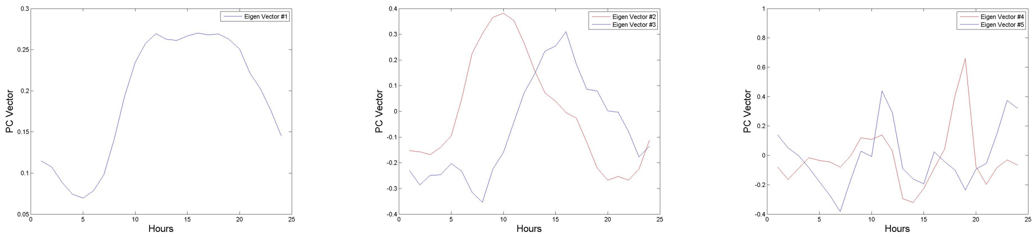


(a) Rush Hour(7-8am) (b) Night(11pm-0am)

Fig. 7: Reconstruction of Vehicle Density Using only 5 Spatial PCs. The x-axis presents the different geographic districts of Beijing, and the y-axis indicates the taxi density.

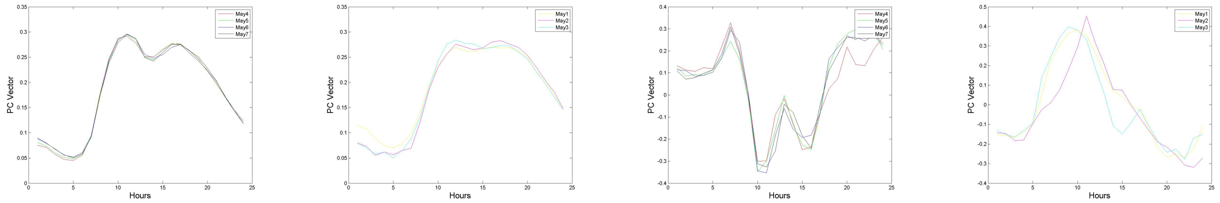
B. Spatial Spectral Structure

Interpretation of PCs. The spatial PCs are organized in a remarkable fashion: As shown in Fig. 11, the first spatial PC represents the mean of taxi density in different districts, in the rank of district population (Fig. 11(a)). As expected, this first spatial PC captures the most important element of statistical variance along the spatial dimension. Interestingly, each of the remaining PCs ((Fig. 11(b)(c)) represents a spike corresponding to a given geographic district, indicating the dominating role of a particular district in each PC. Furthermore, in each spatial PC, we notice there exist negative correlation between different districts, which illustrates the functions of geographic



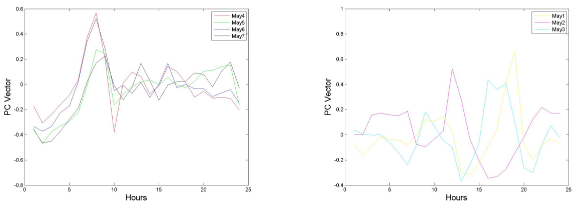
(a) 1st PC (b) 2nd PC & 3rd PC (c) 4th & 5th PC

Fig. 8: The Temporal Principle Components (PCs) of Beijing Trace (May 1). (a) the 1st PC represents the average volume of daily taxi traffic; (b) the 2nd PC and 3rd PC resemble morning traffic and afternoon traffic trend, respectively; (c) the 4th PC and 5th PC represent the morning "rush hour" and after "rush hour".



(a) 1st PC (Weekday) (b) 1st PC (Weekend) (c) 2nd PC (Weekday) (d) 2nd PC (Weekend)

Fig. 9: The Comparison of the 1st and 2nd Principle Components in Temporal Domain, Across a Week, in Beijing Trace. It could seen that the first few of temporal PCs are highly stable among week days or among weekend days.



(a) 4th PC (Weekday) (b) 4th PC (Weekend)

Fig. 10: The Comparison of the 4th Principle Components in Temporal Domain, Across a Week, In Beijing Trace.

districts (e.g., taxis in suburban districts flow into central CBD districts in morning rush hour).

TABLE II: Pearson Correlation of Spatial PCs Between Different Days

PC	Between Weekend	Between Weekday	Between Weekend and Weekday
1st PC	0.9944	0.9987	0.9857
2nd PC	0.7982	0.9670	-0.1116
3rd PC	0.7088	0.8361	-0.0645
4th PC	0.6453	0.7413	0.0244
5th PC	0.4363	0.6353	-0.0460

Stability and Repeatability. Table II shows the stability of top spatial PCs over different combinations (weekday vs. weekday, weekend vs. weekend, and weekday vs. weekend). It is observed that the Pearson correlation of top spatial PCs is high (≥ 0.65) over the weekdays, indicating a consistent spatial spectral structure. Though weaker (Pearson correlation $[0.43, 0.99]$), the spatial spectral structure between weekend is still fairly stable. On the other hand, there is no correlation in these spatial PCs between a weekday and a weekend day, suggesting totally different travel patterns. Fig. 12 further illustrates this point by showing the comparison of the the 1st and 2nd spatial spatial PCs over weekday and weekend.

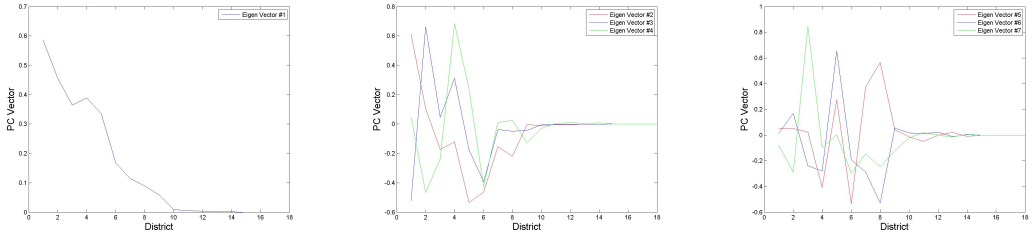
VI. Traffic Anomaly Detection

In this section, we demonstrate how *the stable spectral structure revealed by PCA analysis could be used to detect and localize traffic anomaly events*.

Methodology. The spectral structure is able to isolate traffic events presented in anomalous PCA subspace from the regular events happening in normal PCA subspace. Following PCA anomaly detection literature [18], [19], [20], we partition the traffic matrix $X(p,t)$ at a given time t and location p into two subspaces as

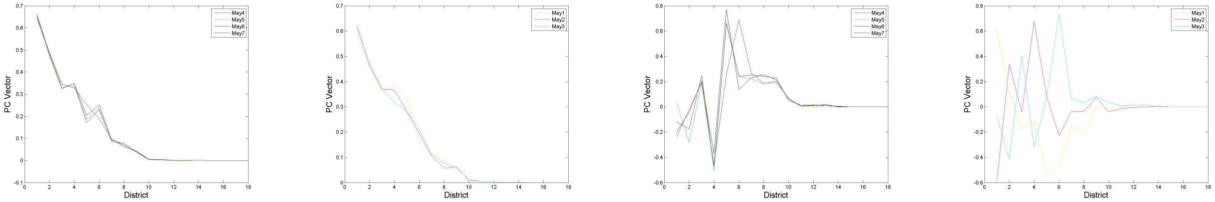
$$\begin{aligned} X(p,t) &= \bar{X}(p,t) + \tilde{X}(p,t) \qquad (5)\\ &= PP^T X(p,t) + (1 - PP^T)X(p,t) \qquad (6) \end{aligned}$$

in which $P = [v_1, v_2, ..., v_r]$ is a truncated matrix representing the normal subspace and it is composed of r-rank principle components obtained via PCA analysis. $\bar{X}(p,t)$ is the regular traffic counts corresponding to normal subspace, and $\tilde{X}(p,t)$ reflects the abnormal traffic events exhibited in anomalous subspace. $\bar{X}(p,t)$ and $\tilde{X}(p,t)$ are obtained by projecting $X(p,t)$ onto the normal and anomalous subspace, respectively. In our study, we choose the Period PCs only as our normal subspace, leaving both spike PCs and Gaussian noise PCs to our anomalous subspace. We argue that this policy is reasonable, because Period PCs captures the vast majority of traffic data variability (Sec. IV), representing the normal state; any major deviation from this normal state is therefore considered an anomaly.



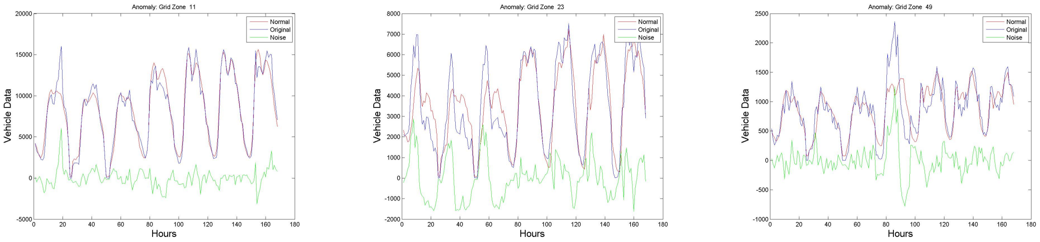
(a) 1st PC (b) 2nd-4th PC (c) 5th-7th PC

Fig. 11: The Spatial Eigen Vectors of Beijing Trace (May 1). The 1st spatial PC represents the mean of vehicle density in each district, and other remaining PCs represent a spike corresponding to its particular district.



(a) 1st PC (Weekday) (b) 1st PC (Weekend) (c) 2nd PC (Weekday) (d) 2nd PC (Weekend)

Fig. 12: The Comparison of the 1st and 2nd Principle Components in Spatial Domain, Across the Week, For Beijing Trace. It could be seen that spatial PCs are fairly stable among week days or among weekend days.



(a) the 11th Grid Zone (b) the 23th Grid Zone (c) the 49th Grid Zone

Fig. 13: Examples of Traffic Density Anomalies of Microscopic Grid Zones in Beijing Across One Week. The traffic density anomaly event is caused by the fact that certain facilities in this zone attract people's visits, e.g., (a) a stadium for a sport event on holiday night, (b) an amusement park during 3 vacation days, and (c) an airport for post-holiday travels.

Experiment Results. We apply the traffic anomaly detection to Beijing trace. To scrutinize the causality effect between holiday schedules and taxi density anomaly events, we partition metropolitan Beijing area into a large number of much smaller grid zones (with each grid zone being 1 km^2). Fig. 13 shows three such examples of taxi density anomaly events. Each example could be explained by the fact that this particular zone has certain facilities that attracts people's visits, e.g., a stadium for a sport event on holiday night (Fig. 13(a)), an amusement park during 3 vacation days (Fig. 13(b)), and an airport for post-holiday travels (Fig. 13(c)).

VII. Related Works

Our study is inspired by pioneering works in two domains.

Spectral Analysis of Transportation Data. Somewhat to our surprise, we found that there are very few research works that attempt to study transportation traffic from a spectral analysis perspective, with the exception of missing data imputation problem. For such missing data imputation problem, spatio-temporal methods (e.g., neighborhood interpolation[8], [9], [10], spline regression [11], [12], [13]) were mainstream solutions [5], [14]; however, transportation researchers began to realize the advantages of spectral analysis in the past decade. BPCA was initially explored to address the traffic data incompleteness problem, which common exists in urban environments due to sensor malfunctions [16]. Later on, a PPCA based solution was shown to perform significantly better than the initial BPCA approach [15], and another PCA variant, Robust PCA [17], was shown to gracefully handle even more challenging scenarios featured with outliers and significant missing data. It is important to note that, unlike prior works narrowly focused on data incompleteness issue, our study has had a more ambitious goal to establish a generic framework, revealing the (hidden) spectral structure of vehicular traffic.

Internet Traffic Engineering. The research work most relevant to our study is Internet traffic matrix analysis, in which Principle Component Analysis and its variants are commonly used mathematical tools. In their seminal paper, Lakhina et al [21] advocated the PCA method for such analysis, and then demonstrated that complex Internet Origin-Destination (OD) matrix could be approximated by a much smaller number of principle components. Realizing that PCA decomposition distinguishes abnormal subspace from normal subspace fairly accurately, they reported that PCA analysis is also an effective method for Internet anomaly detection [22], but they also pointed out that the PCA parameters should be carefully selected to ensure the efficacy of PCA-based anomaly detection method [25]. Several enhancement approaches had been studied to tackle this parameter sensitivity challenge[23]. A distributed PCA solution was proposed to achieve a better trade-off between communication overhead and anomaly detection accuracy [23]. Network anomograph was designed to extend PCA methods to be more generic, covering both spatial domain and temporal domain [24]. To have a more robust performance under challenging scenarios (e.g., large volume of Internet traffic anomalies), it was suggested that the traffic variation matrix – rather than the original traffic matrix – should be used in PCA-based anomaly detection [27]. Finally, discovering that classic PCA method fails to capture temporal correlation, a time-domain Karhunen-Loeve expansion was added into PCA analysis framework [26] to enhance the performance.

Of course, our paper differs from these studies because of the different application domains (Internet traffic vs. vehicle transportation). Another salient difference is that our work examines the geographic structure of vehicular traffic through spectral analysis, in addition to conventional PCA analysis that only focusing on the periodicity (temporal domain) of Internet traffic patterns.

VIII. Conclusion

In this paper, we study the spectral structure of urban vehicular traffic, using spectral analysis tools such as PCA. By analyzing publicly available sets of taxi GPS traces, we find that there indeed exist a stable, low-dimensional spectral structure of urban mobility patterns which is unnoticeable in the conventional spatio-temporal domain. This stable, low-dimensional spectral structure not only significantly simplifies the representation of high-dimension transportation data, but also offers interpretable insights to urban mobility patterns. This is clearly the advantage of spectral analysis over classic spatio-temporal data analysis. Furthermore, leveraging the knowledge obtained through spectral analysis, using concrete examples, we showcase that practical challenges (traffic anomaly) could be tackled.

References

[1] N. cressie, C. K. Wikle, Statistics for Spatio-Temporal Data, Wiley Publishing House.

[2] K. G. Courage, M. Doctor, S. Maddula, and R. Surapaneni, Video image detection for traffic surveillance and control, *Transp. Res. Center, Univ. Florida, Gainesville, FL, USA, Tech. Rep. TD100:FL96- 119, Mar. 1996.* data by video detection for use in transportation planning, *J. Intell. Transp. Syst., vol. 5, no. 4, pp. 343–361, 2000.*

[3] E. Uhlemann, “Autonomous vehicles are connecting… [connected vehicles],” *IEEE Veh. Technol. Mag., vol. 10, no. 2, pp. 22–25, Jun. 2015.*

[4] C. Chen, T. H. Luan, X. Guan, N. Lu, and Y. Liu. (2017). Connected vehicular transportation: Data analytics and traffic-dependent networking. *[Online]. Available: https://arxiv.org/abs/1704.08125*

[5] M. Treibera and D. Helbing, Reconstructing the spatio-temporal traffic dynamics from stationary detector data, *Coop. Transp. Dyn., vol. 1, pp. 3.1–3.24, 2002.*

[6] L. Zhu, F. R. Yu , Y. Wang, B. Ning, and T. Tang, Big Data Analytics in Intelligent Transportation Systems: A Survey, in *IEEE Trans. Intel. Trans. Sys., Vol. 20, No. 1.*

[7] S. Turner, L. Albert, B. Gajewski, and W. Eisele, Archived intelligent transportation system data quality: Preliminary analyses of San Antonio TransGuide data, in *Trans. Res. Rec., no. 1719, pp. 77–84, 2000.*

[8] A. Afifi and R. Elashoff, Missing observationsin multivariate statistics I: Review of the literature, *J. Amer. Stat. Assoc., vol. 61, no. 315, 1966.*

[9] J. L. Schafer, Analysis of Incomplete Multivariate Data, *Monographs on Statistics and Applied Probability, no. 72. London, U.K.: Chapman & Hall, 1997.*

[10] J. Chen and J. Shao, Nearest neighbour imputation for survey data, *J. Off. Stat., vol. 16, no. 2, pp. 113–131, 2000*

[11] J. H. Conklin and B. L. Smith, The use of local lane distribution patterns for the estimation of missing data in transportation management systems, *Trans. Res. Rec., no. 1811, pp. 50–56, 2002.*

[12] D. Ni, J. D. Leonard, II, A. Guin, and C. Feng, Multiple imputation scheme for overcoming the missing values and variability issues in ITS data, *J. Transp. Eng., vol. 131, no. 12, pp. 931–938, Dec. 2005.*

[13] A. Stathopoulos and T. Tsekeris, Methodology for processing archived ITS data for reliability analysis in urban networks, *Proc. Inst. Elect. Eng.—Intell. Transp. Syst., vol. 153, no. 1, pp. 105–112, Mar. 2006.*

[14] H. S. Zhang, Y. Zhang, Z. H. Li, and D. Hu, Spatial–temporal traffic data analysis based on global data management using MAS, *IEEE Trans. Intell. Transp. Syst., vol. 5, no. 4, pp. 268–275, Dec. 2004.*

[15] L. Qu, L. Li, Y. Zhang, J. Hu. PPCA-Based Missing Data Imputation for Traffic Flow Volume: A Systematical Approach, in *IEEE Transaction on Intelligent Transporation Systems, Vol.10, No. 3.*

[16] L. Qu, J. Hu, and Y. Zhang, A flow volumes data compression approach for traffic network based on principal component analysis, in *Proc. IEEE Intell. Transp. Syst. Conf., Seattle, WA, 2007, pp. 125–130.*

[17] S. Serneels, T. Verdonck, Principle component analysis for data containing outliers and missing elements, in *Elsservier Computationals Statistics and Data Analytics, Vol. 52 (2008).*

[18] R. Dunia and S. J. Qin. Multi-dimensional Fault Diagnosis Using a Subspace Approach. In *American Control Conference*, 1997.

[19] R. Dunia and S. J. Qin. A Subspace Approach to Multidimensional Fault Identification and Reconstruction. In *American Institute of Chemical Engineers (AIChE) Journal*, pages 1813–1831, 1998.

[20] J. E. Jackson and G. S. Mudholkar. Control Procedures for Residuals Associated with Principal Component Analysis. in *Technometrics*, pages 341–349, 1979.

[21] A. Lakhina, K. Papagiannaki, M. Crovella, C. Diot, E. D. Kolaczyk, and N. Taft. Structural analysis of network traffic flows. in *SIGMETRICS Perform. Eval. Rev. 32, 1 (June 2004), 61-72.*

[22] A.Lakhina, M.Crovella, C.Diot. Diagnosing network-wide traffic anomalies. *SIGCOMM Comput.Commun.Rev.34, 4(August2004), 219-230.*

[23] L Huang, X. Nguyen, M. Garofalakis, M. Jordon, A. Joseph and N. Taft. In-network PCA and anomaly detection. In *Proceedings of Neural Information Processing Systems (NIPS) 2006, December 2006.*

[24] Y.Zhang, Z.Ge, A.Greenberg, M.Roughan Network anomography. In *Proceedings of the 5th ACMSIGCOMM conference on Internet Measurement (IMC '05). USENIX Association, Berkeley, CA, USA, 317-330.*

[25] H. Ringberg, A. Soule, J. Rexford, and C. Diot. Sensitivity of PCA for traffic anomaly detection. In *SIGMETRICS Perform. Eval. Rev. 35, 1 (June 2007), 109-120.*

[26] D. Brauckhoff, K. Salamatian, M. May, Applying PCA for Traffic ANomaly Detection: Problem and Solution, in *Proceeding of IEEE Infocom 2009.*

[27] Y. Ohsita, S. Ata, and M. Murata. Identification of Attack Nodes from Traffic Matrix Estimation. in *IEICE Transactions on Communications, Vol.E90-B, No.10 (Oct 2007). 2854-2864.*

[28] R. Molina-Masegosa, J. Gozalvez, LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications, in *IEEE Vehicular Technology Magazine*, 12(4), 30–39, 2017.