A functional data approach to missing value imputation and outlier detection for traffic flow data

YI-CHEN ZHANG
Department of Statistics and Probability
zhang318@stt.msu.edu

Abstract

Missing values and outlier identification problems are frequently encountered in traffic monitoring data. We approach these problems by sampling the daily traffic flow rate trajectories from random functions and taking advantage of the data features using functional data analysis. We propose to use the conditional expectation approach to functional principal component analysis (FPCA) to impute missing values. Following the FPCA approach, the functional principal component scores can be applied to the functional bagplot and functional highest density region (HDR) boxplot, which makes outlier detection possible for incomplete functional data.

Keywords: functional data; functional principal component analysis; intelligent transportation system; traffic flow rate; vehicle detector.

1. INTRODUCTION

Traffic monitoring data provide valuable information for highway planning and traffic surveillance and control purposes. Real-time traffic monitoring data provide essential information for traffic surveillance and control in Intelligent Transportation Systems (ITS). Applications of the traffic monitoring data require complete and reliable data. These data can be recorded automatically by various types of vehicle loop detectors, which are usually installed in a planned road with regular intervals. Since loop detectors operate in a rough environment, missing data problems are inevitable due to detector malfunctions or package loss during transmission. Temporary detector malfunctions that result in loss of data are quite common.

One way to deal with missing values is to eliminate samples with missing values from the original dataset, yet the reduced dataset may lead to biased analysis results. Another approach is to reconstruct missing entries based on the recorded dataset; however, distinct imputation methods have their own advantages and disadvantages with different imputation performances depending on data availability scenarios. Each method can lead to different imputing results. Just like normal data collection or analysis procedures, missing data should be an important consideration in designing traffic data archiving or analysis systems for the purposes of highway planning and traffic surveillance and control, especially for applications in ITS. In addition, outlier detection is another important issue in investigating traffic data. In addition to detecting temporal outliers in terms of magnitude outliers in time, identifying unusual patterns of trajectories (i.e., shape outlier) is also important. Analyzing the causes on the shape outliers can provide different and useful information for further applications to traffic management.

Resources on missing data in general can be found in Allison (2003) and Schafer and Graham (2002). Methods specifically discussed for traffic flow data have attracted significant attention. These include the Kalman filter method (Dailey, 1993), time series modeling (Nihan, 1997), historical (neighboring) imputation (Chen and Shao, 2000), the lane distribution method (Conklin and Smith, 2002), spline regression imputation methods (Chen et al., 2003) and genetically designed modeling (Zhong et al., 2004). More recently, Ni et al. (2005) proposed a multiple imputation scheme for imputing missing values. Qu et al. (2009) proposed Probabilistic Principal Component Analysis (PPCA) and Bayesian Principal Component Analysis (BPCA) imputation algorithms and compared their performance with some conventional methods from the literature. Although numerous multivariate analysis methods have been developed to deal with missing values, to the best of our knowledge, functional data approaches to take advantage of functional data features have not yet been discussed in relation to imputing missing values for longitudinal or functional data.

As for the methods used for outlier detection, like many statistical analysis procedures, one of the first steps toward obtaining a coherent analysis is the detection of outlying observations. While outlier detection of multivariate data has been developed over several decades, outlier detection of functional data has only been discussed in recent years. Identification of abnormal or unusual patterns of trajectories that deviate significantly from other observations in a homogeneous group can improve the quality of observations and for further research. Abnormal data may adversely lead to model misspecification, biased parameter estimation and incorrect results. Methods of outlier detection of functional data that are found in the literature include the use of robust principal component analysis (Hyndman and Ullah, 2007), the successive likelihood ratio test and smoothed bootstrapping (Febrero et al., 2007), singular value decomposition

plots (Zhang et al., 2007), rainbow plots, bagplots and boxplots for functional data (Hyndman and Shang, 2010) and functional boxplots (Sun and Genton, 2011).

In this study, we consider a functional data analysis (FDA) approach, where daily traffic flow trajectories are treated as functional data that are sampled from random functions. FDA was introduced nearly two decades ago and various statistical methods for FDA have been extensively developed. Overviews of the methodological foundations of FDA can be found in (Ramsay and Silverman, 2005; Ferraty and Vieu, 2006), as well as the review articles (Rice, 2004; Müller, 2005). One of the most popular approaches to FDA is Functional Principal Component Analysis (FPCA), which was motivated by the Karhunen-Lóeve expansion for stochastic processes. In this paper, we propose FPCA based methods to handle incompleteness of traffic flow data. Moreover, following the missing data imputation method, we provide two visualized outlier detection methods based on functional principal component (FPC) scores: the modified functional bagplot and the modified functional highest density region (HDR) boxplot, both of which are graphical tools aimed at identifying outlying curves of functional data.

This article is organized as follows. In Section 2, we describe the patterns of missing data and of outlying curves from the functional data point of view. Section 3 introduces the theoretical background of FPCA and the approach to imputing missing values via FPCA techniques, followed by the modified outlier detection methods. A real data analysis carried out by our proposed FPCA approach is presented in Section 4. Concluding remarks and discussions are provided in Section 5.

2. Missing Data and Outliers in Traffic Flow Trajectories

There are many problems encountered in analyzing traffic flow data, one of which is the quality of the data. Although, traffic flow is automatically recorded by dual loop detectors, there are some opportunities for data corruption, such as short-term software or hardware malfunctions, maintenance operations and detector construction. The resultant effects of these are outliers and discontinuities or gaps in the data record, both of which create severe obstacles in modeling and identification of the underlying process. Without knowing the structure and details of missing data and outliers, problems may arise when analysis is performed. Before any such analysis, it is useful to perform a review of the dataset in order to fill missing gaps and remove identified outliers.

2.1 Patterns of missing value

Missing data can be random in nature and are sometimes caused by a detector that does not deliver measurement values, a fault in the measurement tools or even human error. Depending on the measurement facility, missing values can appear as a blank, zero, negative, or not a number. Therefore, missing values are often simple to detect in a recorded dataset. For the traffic flow data we simply categorize the missing patterns as follows:

- Point Missing (PM): The missing points are completely independent of the observed and unobserved values. The missing points are isolated, grouped or randomly scattered. See Figure 1(a).

- Interval Missing (IM): In terms of functional data, the observed data are curves instead of points. Hence, a missing interval means an unobserved interval rather than some unobserved points in a small group. The missing intervals often occur randomly. See Figure 1(b).

- Mixed PM/IM: The missing patterns can be PM or IM. See Figure 1(c).

Data incompleteness is a troubling feature of many datasets and missing values are a serious problem as they may distort the properties of the data. Although there may be different patterns of missing values, the imputation method we propose is based on the partially observed trajectories and is not affected by the missing patterns.

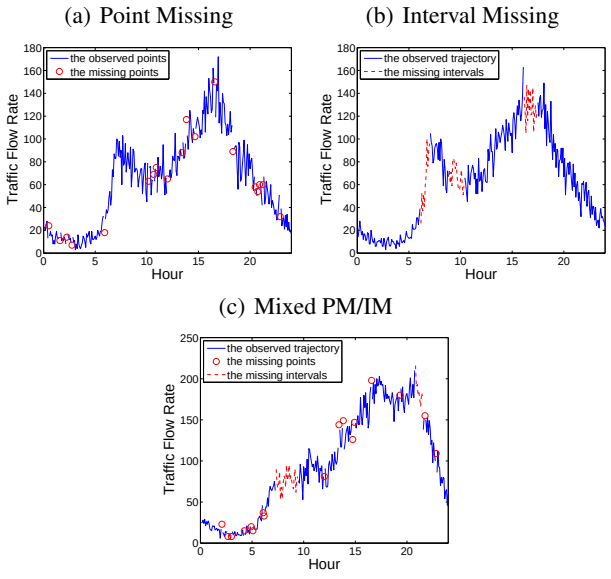

Figure 1: Typical missing patterns of traffic flow data. (a) Point Missing (PM): The circles are missing points. (b) Interval Missing (IM): The dotted lines are missing intervals. (c) Mixed PM/IM.

2.2 Patterns of outlying curve

Outlier detection is a prerequisite in many data applications. There are several methods for outlier detection that can be distinguished as univariate/multivariate techniques and parametric/non-parametric procedures. For instance, the Mahalanobis distance is a well-known criterion that depends on estimated parameters of the multivariate distribution. Although there are many outlier detection methods for multivariate data, very few of them are for functional data. Defining an outlier or a contamination with a sample of curves is itself a tricky problem. Detecting outlying curves is a challenging task and mistakes or oversights in this area can have serious effects on statistical analysis, including biasing the results.

Following Hyndman and Shang (2010), there are two types of outliers, magnitude outliers and shape outliers. In general, magnitude outliers are distant from the mean and shape outliers have a pattern that is different from the other curves, e.g., see Figures 2(a) and 2(b), respectively. In practice, outlying curves may exhibit a combination of these features. When analyzing functional data, outliers can greatly affect estimates in many ways, including skewing the summary statistics and distorting the statistical modeling. Further research based on such models and summaries can result in potentially serious failure due to previously undetected errors. Thus, identifying outliers can be vital. A good methodology for trying to identify outlying curves should be able to cope with all types of outliers.

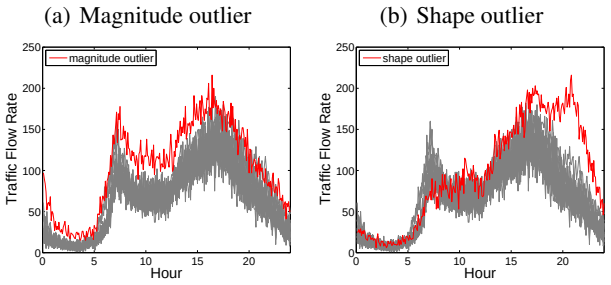

Figure 2: Typical outliers of traffic flow rate trajectories. (a) Magnitude outlier; (b) Shape outlier.

3. FUNCTIONAL PRINCIPAL COMPONENTS ANALYSIS

Most functional data approaches are nonparametric due to the data features, which impose minimal assumptions on the data that overcome the limitations in parametric modeling. In this section, we make use of functional principal component analysis (FPCA) techniques to impute missing values of functional data.

3.1 Functional principal component models

We adopt the notion that each daily traffic flow trajectory is a realization of a random function. Let X denote the random function for the daily traffic flow trajectory which is assumed to be a smooth random function. We further assume that the random function X has an unknown mean function $EX(t) = \mu(t)$ and covariance function $\text{cov}(X(s), X(t)) = G(s,t)$, $s, t \in \mathcal{T}$, where $\mathcal{T} = [0, T]$ is a bounded and closed time interval in the L^2 space. Each function in the L^2 space can be expressed in terms of basis functions generated from the space. Here we assume that G has an orthogonal expansion in L^2, that is, $G(s,t) = \sum_{k=1}^{\infty} \lambda_k \phi_k(s)\phi_k(t)$, where $\{\lambda_k\}$ is a set of eigenvalues with the order $\lambda_1 \geq \lambda_2 \geq \ldots$ and $\{\phi_k\}$ is the associated set of eigenfunctions that form a set of basis in L^2. A random trajectory from the traffic flow then has the following Karhunen-Loève representation:

$$X_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t) \tag{1}$$

where $\xi_{ik} = \langle X_i - \mu, \phi_k \rangle$ is a random coefficient, projecting $(X_i - \mu)$ in the direction of the kth eigenfunction ϕ_k, with a mean of zero and variance λ_k. The representation of X_i in (1) contains an overall mean function μ, which describes the main trend in t of all trajectories, as well as a sequence of basis function ϕ_k multiplied by random coefficients ξ_{ik}, which characterizes the individual trajectories varying in the functional space. This representation provides a useful technique for of reducing the dimension of random functions since the random element can often be well approximated by the first few leading components.

3.2 Estimation of functional principal component model

In practice, the random function X_i is often contaminated with measurement or experimental errors. The observations of the ith data object, $i = 1, \ldots, n$, with m_i observations observed at t_{ij} for all t_{ij} in $\mathcal{T}$ and $j = 1, \ldots, m_i$, can be represented as

$$\begin{aligned} Y_i(t_{ij}) &= X_i(t_{ij}) + \varepsilon_{ij} \\ &= \mu(t_{ij}) + \sum_{k=1}^{\infty} \xi_{ik}\phi_k(t_{ij}) + \varepsilon_{ij}, \end{aligned} \tag{2}$$

where X_i is the random function described in (1), and the additional measurement random errors ε_{ij} are assumed to be uncorrelated with each other and are independent of the random coefficients ξ_{ik}, with $E(\varepsilon_{ij}) = 0$ and $\text{var}(\varepsilon_{ij}) = \sigma^2$. The covariances of Y are

$$\text{cov}(Y(t_{ij}), Y(t_{il})) = \text{cov}(X(t_{ij}), X(t_{il})) + \sigma^2\delta_{jl} \tag{3}$$

where δ_{jl} is 1 if $j = l$ and 0 otherwise. The covariance expression (3) indicates that the measurement error contributes

additional variance on the diagonal of the covariance surface. To obtain the corresponding function estimate in (1), we must estimate the model component functions μ and ϕ_k. We apply the locally weighted least squares smoothing method on the pooled data from all trajectories for the estimated mean function μ. The smoothing parameters can be chosen by various methods, including cross-validation (Rice and Silverman, 1991) or generalized cross-validation (Fan and Gijbels, 1996). We adopt the techniques proposed in Yao et al. (2003) and smooth the empirical covariances to obtain the estimate of G. The estimated eigenvalue $\hat{\lambda}_k$ and $\hat{\phi}_k$ can be derived numerically by applying the eigen-decomposition procedure from the smoothed estimate of the covariance function. Finally, the fitted covariance surface is obtained by $\widehat{G}(s,t) = \sum_{k=1}^{L} \hat{\lambda}_k \hat{\phi}_k(s)\hat{\phi}_k(t)$, where L denotes the number of components included in the Karhunen-Loève representation above (1). Here, we choose L to be the smallest value such that the first L components explains at least $\tau_\lambda \times 100\%$ of total variance, such that

$$L = \min\left\{ L \geq 1 : \frac{\sum_{k=1}^{L} \hat{\lambda}_k}{\sum_{k=1}^{M} \hat{\lambda}_k} \geq \tau_\lambda \right\}, \tag{4}$$

where M is the largest number of components with $\hat{\lambda}_k > 0$ and τ_λ is a predetermined threshold value, $0 \leq \tau_\lambda \leq 1$. Setting τ_λ equals 0.9 or higher works reasonably well in our data applications.

The simplest method for estimating functional principal component scores is to obtain the estimate of ξ_{ik} by $\hat{\xi}_{ik} = \int (X_i(t) - \hat{\mu}(t))\hat{\phi}_k(t)\, dt$ using numeric approximation. However, this integral approximation method encounters difficulties when there are many missing entries or only a few repeated observations available. In addition, since the observation Y_is are contaminated with measurement errors, estimating ξ_{ik} by substituting Y_i for X_i may lead to biased functional principal component scores. To overcome these difficulties, we adopt the approach of Yao et al. (2005) in relation to the conditional expectation by assuming that in (2), ξ_{ik} and ε_{ij} are jointly Gaussian. Let $\boldsymbol{Y}_i = (Y_i(t_{i1}), \ldots, Y_i(t_{im_i}))^T$, where m_i is the number of available observations for the ith trajectory. Let $\boldsymbol{\phi}_{ik}$ be the vector of the values of the kth eigenfunction, $\boldsymbol{\phi}_{ik} = (\phi_k(t_{i1}), \ldots, \phi_k(t_{im_i}))^T$, $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}$ be the covariance matrix of $\boldsymbol{Y}_i$, and $\boldsymbol{\mu}_i = (\mu(t_{i1}), \ldots, \mu(t_{im_i}))^T$. Under the assumption that the principal components ξ_{ik} and error term ε_{ij} are jointly Gaussian, the conditional principal components are

$$E(\xi_{ik}|\boldsymbol{Y}_i) = \lambda_k \boldsymbol{\phi}_{ik}^T \boldsymbol{\Sigma}_{\boldsymbol{Y}_i}^{-1}(\boldsymbol{Y}_i - \boldsymbol{\mu}_i). \tag{5}$$

The estimated conditional principal components in (5) are then obtained by substituting the corresponding estimates, giving

$$\hat{\xi}_{ik} = \hat{\lambda}_k \hat{\boldsymbol{\phi}}_{ik}^T \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i}^{-1}(\boldsymbol{Y}_i - \hat{\boldsymbol{\mu}}_i), \tag{6}$$

where $\hat{\boldsymbol{\phi}}_{ik} = (\hat{\phi}_k(t_{i1}), \ldots, \hat{\phi}_k(t_{im_i}))$ is the estimate of $\boldsymbol{\phi}_{ik}$ and $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{Y}_i} = \{\widehat{G}(t_{ij}, t_{il}) + \hat{\sigma}^2 \delta_{ij}\}_{1 \leq j,l \leq m_i}$ is the estimate of $\boldsymbol{\Sigma}_{\boldsymbol{Y}_i}$. Note that $\widehat{G}(t_{ij}, t_{il})$ and $\hat{\sigma}^2$ the estimates of $G(t_{ij}, t_{il})$ and σ^2 and δ_{ij} is the Kronecker delta. More details about the conditional expectation approach can be found in Yao et al. (2005).

3.3 Missing value imputation by functional principal component models

In view of (6), this functional principal component score estimate is applicable to situations where there are missing values and, thus, inspires our missing value imputation method. Based on the estimated model components $\hat{\mu}$, $\hat{\phi}_k$ and $\{\hat{\xi}_{ik}\}_{k=1,\ldots,L}$ for all i as described in the previous subsection, the predicted functions for all of the ith data object are then given by

$$\hat{Y}_i(t) = \hat{\mu}(t) + \sum_{k=1}^{L} \hat{\xi}_{ik}\hat{\phi}_k(t). \tag{7}$$

Model (7) subject to the ith data object holds for all t in the entire time domain $\mathcal{T}$. Therefore, the model fits can be used to impute missing values. For a fixed i, if the observations at times $\{t_{ij}\}$ are missing, then missing entries $Y_i(t_{ij})$ can be imputed by the predicted values $\hat{Y}_i(t_{ij})$. Notice that the predicted trajectories $\hat{Y}_i$ include the components of the smoothed mean function and a linear combination of the eigenfunctions, which recover individual trajectories from noise measurements. Additionally, the imputation errors depend on the model complexity while the number of functional principal components are determined by the fraction of variance explained (FVE). We note that (4) provides a natural method to determine the number of principal components, and we will investigate the effect of the number of components on the imputed missing values.

3.4 Visualization tools for outlier detection

A visualization tool can be useful to detect abnormal random trajectories for functional data when data are contaminated with outlying curves. Two graphical tools were proposed in Hyndman and Shang (2010), functional bagplot and functional highest density region (HDR) boxplot, to detect outliers. Both are based on the first two principal component scores. For this purpose, Croux and Ruiz-Gazen's (2005) robust principal component estimate algorithm is applied with a form of projection pursuit because the principal component decomposition may be sensitive to outliers. In addition, this algorithm is more resistant to outliers when the measurement matrix contains outliers. However, this method does not taking missing values into account; besides, some difficulties can arise when the sample covariance matrix has several hundred or several thousand dimensionalities. Computing such sample covariance itself is very costly. Furthermore, the proper way of dealing with an incomplete dataset

is not clear; particularly as the projection pursuit algorithm requires a complete dataset to project the data onto a lower-dimensional space such that a robust measure of variance of the projected data will be maximized.

To overcome the aforementioned difficulties, instead of using the robust principal component scores, we introduce functional principal component scores based on conditional expectation in a functional bagplot and functional HDR boxplot. Functional principal component scores can capture much of the information inherent in functional data since the covariance surface has smoothed out some outlying features and noise from the measurement errors. We call these two modified outlier detection tools 'modified functional bagplot' and 'modified functional HDR boxplot'. Details about the functional bagplot and the functional HDR boxplot are discussed in Hyndman and Shang (2010).

4. DATA APPLICATIONS

We implement the proposed missing value imputation and outlier detection methods for traffic flow data collected by a dual loop vehicle detector located at 28.45K northbound on National Highway No. 5 in Taiwan, which is is near the entrance of Shea-San Tunnel at 28.11K northbound. The traffic flow rates were collected on a 5-min interval from April 1 to April 30 in 2009. National Highway No. 5 is the major road northbound from Yilan County to Taipei. Yilan County is located nearby Taipei and many people living in Taipei like to go to Yilan County during weekends and holidays for their recreational trips. Therefore, numerous trips northbound on National Highway No. 5 are recreational trips coming back from Yilan County to Taipei, especially starting from the afternoon till evening during weekends and holidays. As shown in Figure 3(a) for the observed traffic flow rate trajectories, the peak hours occurred between 14:00 and 21:00. The data set consists of a sample of 30 functional observations, among which 22 were weekdays and 8 were weekends (including Chinese Tomb-Sweeping holiday on April 5), and each sample contains 288 data observations. Although the data were automatically recorded by the dual loop vehicle detector, there were some missing entries caused by the malfunction of detector, losing packages during transmissions or other reasons.

4.1 Functional principal component analysis

Observing that the traffic flow patterns are distinct on weekends (including holidays) and weekdays, we separate the functional principal component analysis in these two groups. The oberved trajectories and the estimated mean function are displayed in Figures 3(a) and 4(a). The estimated mean functions indicate the peak hours on weekends occur from 14:00 to 20:00, while there two peaks on weekdays, one around 8:00 with lower flow rates and smaller variability and the other around 17:00 with relatively high flow rates and variability. The estimated auto-covariance functions are shown in Figures 3(b) and 4(b). On weekdays the first peak occurs around from 07:00 to 09:00, which is on-work state, and the second peak occurs from 16:00 to 18:00, which is off-work state. The time from 09:00 to 16:00 shows a regular state. The variability is more complicated on weekends with high variability during peak hours. This smoothed covariance surface reveals the structure of the underlying process, which would be difficult for modeling using traditional parametric approaches. In addition, the eigenfunctions from the decomposition of the estimated covariance are shown in Figures 3(d) and 4(d). The number of components were determined by setting $\pi_\lambda = 0.9$. The for leading principal component accounts for 65.81%, 15.20% and 11.03% of total variation for holidays, where the first eigenfunction reflects the overall variability in the peak-hour period. In contrast the two leading components explains 55.01% and 36.42% of total variations in weekdays, where the first and the second eigenfunctions contrast variability between early and late times.

Figure 5 displays samples of observed daily traffic flow trajectories, along with the predicted functions and the imputed missing values of different missing patterns. The imputed results appear reasonable. Particularly, the method can catch curvature pattern in interval missing as shown in Figure 5(c).

4.2 Outlier detection results

Detecting extreme traffic flow trajectories through visual inspection can be difficult due to large noise and missing values. We perform outlier detection by applying the modified functional bagplot and the modified functional HDR boxplot, both of which use the functional principle component scores based on the conditional expectation approach.

The outlier detection results based on the modified functional bagplot are shown in Figure 6. Figure 6(a) displays the modified bivariate bagplot, where the red star marks the Tukey median of the bivariate functional principal component scores, the dark gray region displays the 50% bag and the light gray region shows the 95% fence. The points at April 4 and April 5, outside the fence are identified as outliers. The modified functional bagplot is shown in Figure 6(b), where the solid black curve (median curve) corresponds to the median point (red star) and the similar shaded dark and light gray region corresponds to the bag and fence in the modified functional bagplot. The outlying curve at April 4 and April 5 are highlighted in green and red. In this dataset, the suspected functional outlier dated April 5 is Chinese Tomb-Sweeping holiday. Figure 6(c) illustrates the modified bivariate HDR boxplot using the setting of $\alpha = 0.15$. We note that when using the setting of $\alpha = 0.01$ to 0.1 April 5 is the only identified outlier. Figure 6(d) dis-

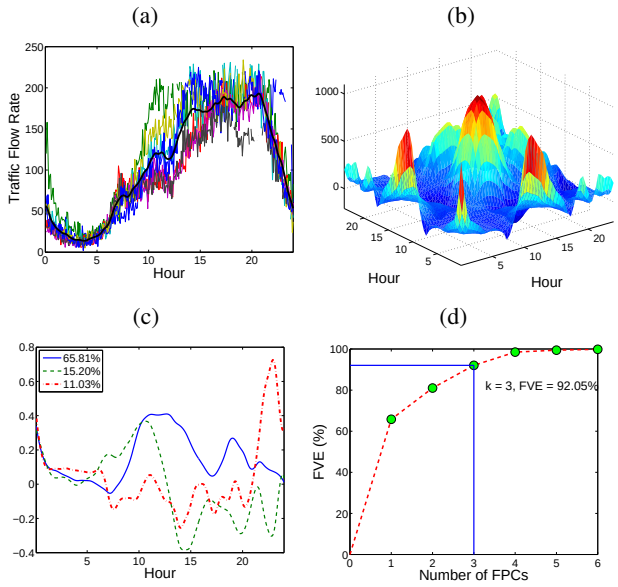

Figure 3: (a) Daily traffic flow trajectories superimposed on the estimated mean function, (b) the estimated covariance function, (c) the estimated eigenfunctions, and (d) the cumulative fraction of total variance explained by the leading FPCs for weekends (including holidays).

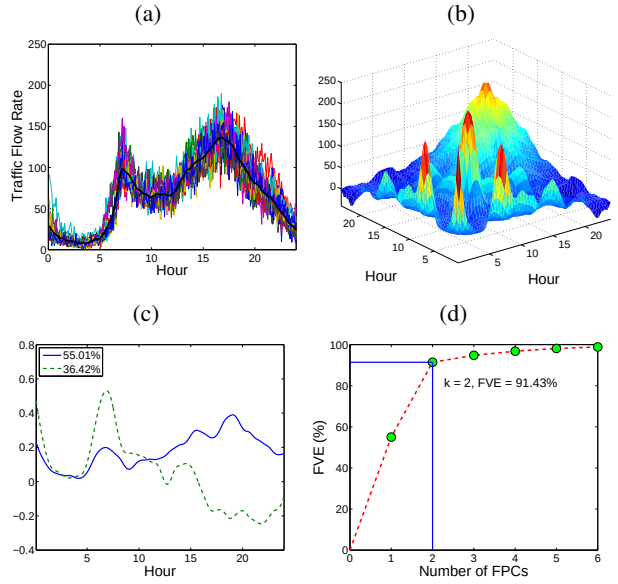

Figure 4: (a) Daily traffic flow trajectories superimposed on the estimated mean function, (b) the estimated covariance function, (c) the estimated eigenfunctions, and (d) the cumulative fraction of total variance explained by the leading FPCs for weekdays.

plays the corresponding modified functional HDR boxplots.

More descriptions on the modified bivariate HDR box-

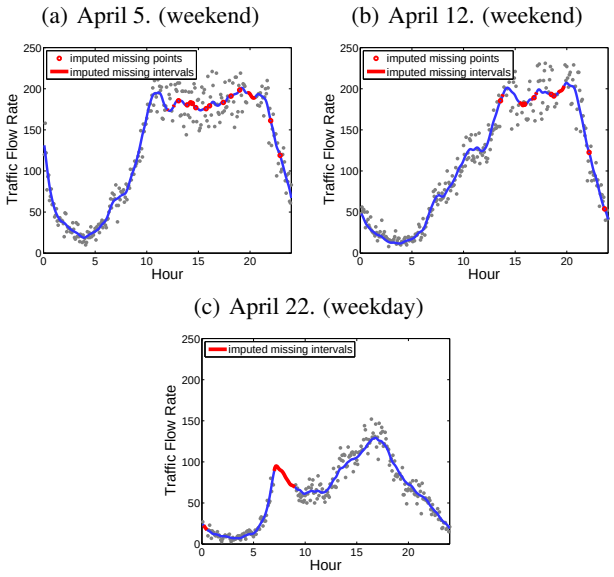

Figure 5: Three samples of daily traffic flow rate trajectories with the observations (in gray), the predicted trajectories (in blue) and the imputed missing values (in red).

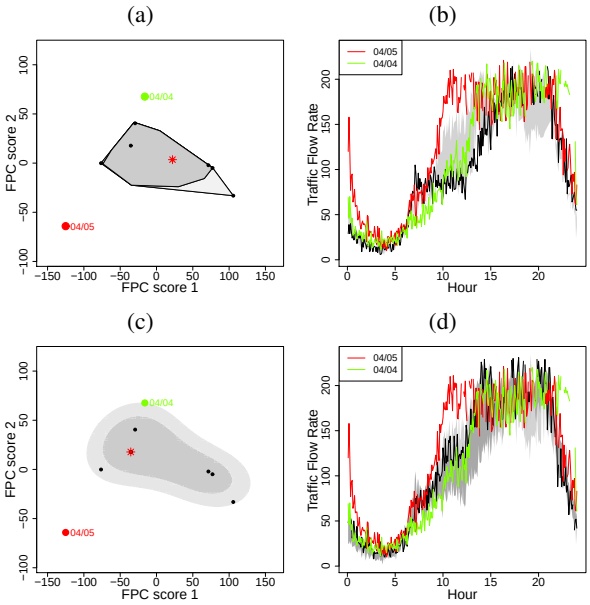

Figure 6: (a) The modified bivariate bagplot, (b) the modified functional bagplot, (c) the modified bivariate HDR boxplot (with $\alpha = 0.15$), and (d) The modified functional HDR boxplot for outlier detection on weekends.

plot and functional boxplot will be followed. April 5 is on Sunday and it is also the Chinese Tomb-Sweeping holiday. It is a special day for people returning back to their hometowns for getting together with their families. In addition to

recreational trips, there are many back-to-hometown trips. Therefore, it is understand that the traffic flow pattern on April 4 and April 5 are quite different from other weekends as illustrated in Figure 6.

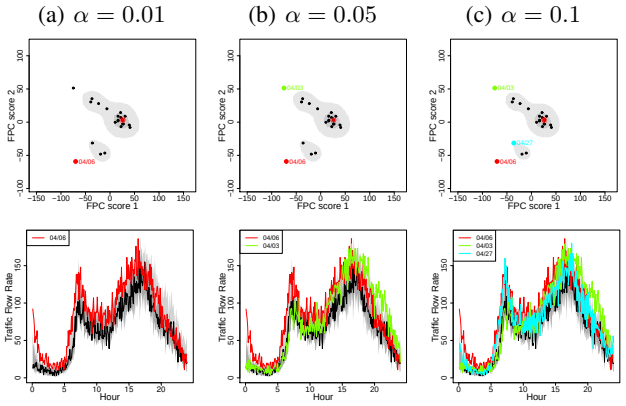

Figure 7: The modified bivariate HDR boxplots with $\alpha = 0.01$, $\alpha = 0.05$ and $\alpha = 0.1$. Top panels: the modified bivariate HDR boxplot. Bottom panels: the modified functional HDR boxplot.

For weekdays, no outliers are detected based on the modified functional bagplot approach. To use the functional HDR boxplots, a prespecified coverage probability of the outlying region is needed. We use three coverage probabilities 99%, 95%, and 90%, corresponding to the settings of $\alpha = 0.01, 0.05$, and 0.1, to perform the outlier detection procedure for the traffic flow data. The top panels of Figures 7 illustrate the modified bivariate HDR boxplots in which the red star marks the mode of the bivariate functional principal component scores, the darker gray regions display the 50% HDR, and the lighter gray regions display the 99%, 95% and 90% HDR, respectively. The points outside the light gray regions are identified as outliers. The bottom panels of Figures 7 display the corresponding modified functional HDR boxplots, where the black curves correspond to the mode of bivariate functional principal component scores, and the shaded dark and light regions correspond to the regions in the modified bivariate HDR boxplots. The modified functional HDR boxplots with $\alpha = 0.01, 0.05$, and 0.1 detect one, two and three outliers in the order of April 6, 3 and 27, which are highlighted in red, blue, and green, respectively. The capability of identifying outliers using the functional HDR boxplots highly depends on the pre-specified α. The results show that more outliers are detected with larger values of α and the strength of potentially flagged outliers can be identified by varying the values of α. While the curve corresponding to April 27 is very close to the boundary of the 90% region, the identified outliers on April 3 (Friday) and April 6 (Monday) are both around the April 5 Chinese Tomb-Sweeping holiday, which gives reasonable interpretation. Based on the outlier detection results, we found that traffic flow patterns on the special holiday of April 5, and the days before and after the holiday are different from the general weekend or weekday patterns. Traffic control strategies for such holidays require separate considerations for the weekends and the weekdays.

5. Concluding Remarks

In this study, we proposed a nonparametric functional data approach to missing value imputation and outlier detection for functional data. Our method takes advantage of the functional data features that can be expanded by the functional principal component models consisting of the mean function and a stochastic component to catch individual variation. Moreover, a modified version of the functional bagplot and the functional HDR boxplot that applies functional principal component scores was proposed. One of the advantages of the proposed approach is that it can be used even for incomplete or irregularly collected functional data. Although motivated by traffic flow data, the proposed methodology is widely applicable to data that are repeatedly measured over a period of time.

References

[1] Allison, P.D. (2001) Missing data. Thousand Oaks, CA: Sage.

[2] Bishop, C.M. (1999) Bayesian PCA. Advances in Neural Information Processing Systems, 11, 382-388.

[3] Chen, C., Kwon, J., Rice, J., Skabardonis, A., and Varaiya, P. (2003) Detecting errors and imputing missing data for single-loop surveillance system. Transportation Research Record: Journal of the Transportation Research Board, 1855, 160-167.

[4] Chen, J. and Shao, J. (2000) Nearest nieghbor imputation for survey data. Journal of Official Statistics, 16(2), 113-131.

[5] Chiou, J.M. and Müller, H.G. (2009) Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. Journal of the American Statistical Association, 104(486), 572-585.

[6] Conklin, J.H. and Smith, B.L. (2002) Use of local lane distribution patterns for the estimation of missing data values from traffic monitoring system. Transportation Research Record: Journal of the Transportation Research Board, 1811, 50-56.

[7] Croux, C. and Ruiz-Gazen, A. (2005) High break down estimators for principal components: The projection-pursuit approach revisited. Journal of Multivariate Analysis, 95(1), 206-226.

[8] Dailey, D.J. (1993) Improved error detection for inductive loop sensors. Report No. WA-RD 3001. Washington State Department of Transportation.

[9] Fan, J. and Gijbels, I. (1996) Local polynomial modelling and its application. London: Chapman and Hall.

[10] Febrero, M., Galeano, P., and González-Manteiga, W. (2007) A functional analysis of NOx levels: Location and scale estimation and outlier detection. Computational Statistics, 22(3), 411-427.

[11] Ferraty, F. and Vieu, P. (2006) Nonparametric Functional Data Analysis: Theory and Practice. New York: Springer.

[12] Hyndman, R.J. and Ullah, M.S. (2007) Robust forecasting of mortality and fertility rates: A functional data approach. Computational Statistics and Data Analysis, 51(10), 4942-4956.

[13] Hyndman, R.J. and Shang, H.L. (2010) Rainbow plots, bagplots, and boxplots for functional data. Journal of Computational and Graphical Statistics, 19(1), 29-45.

[14] Müller, H.G. (2005) Functional modeling and classification of longitudinal data. Scandinavian Journal of Statistics, 32(2), 223-240.

[15] Müller, H.G., Chiou, J.M., and Leng, X. (2008) Inferring gene expression dynamics via functional regression analysis. BMC Bioinformatics, 9(60).

[16] Nakai, M. and Ke, W. (2011) Review of the methods for handling missing data in longitudinal data analysis. International Journal of Mathematical Analysis, 5(1), 1-13.

[17] Ni, D., Leonard, J.D., Guin, A., and Feng, C. (2005) Multiple imputation scheme for overcoming the missing values and variability issues in ITS data. Journal of Transportation Engineering, 131(12), 931-938.

[18] Nihan, N. (1997) Aid to determining freeway metering rates and detecting loop errors. Journal of Transportation Engineering, 123(6), 454-458.

[19] Oba, S., Sato, M.A., Takemasa, I., Monden, M., Matsubara, K.I., and Ishii, S. (2003) A bayesian missing value estimation method for gene expression profile data. Bioinformatics, 19(16), 2088-2096.

[20] Qu, L., Li, L., Zhang, Y., and Hu, J. (2009) PPCA-based missing data imputation for traffic flow volume: A systematical approach. IEEE Transactions on Intelligent Transportation System, 10(3), 512-522.

[21] Ramsay, J.O. and Silverman, B.W. (2005) Functional Data Analysis. 2nd ed. New York: Springer.

[22] Rice, J.A. and Silverman, B.W. (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society Series B, 53(1), 233-243.

[23] Rice, J.A. (2004) Functional and longitudinal data analysis: Perspectives on smoothing. Statistica Sinica, 14(3), 631-647.

[24] Rubin, D.B. (1987) Multiple imputation for nonresponse in surveys. New York: Wiely.

[25] Schafer, J.L. and Graham, J.W. (2002) Missing data: Our view of the state of the art. Psychological Methods, 7(2), 147-177.

[26] Sun, Y. and Genton, M.G. (2011) Functional boxplots. Journal of Computational and Graphical Statistics, 20(2), 316-334.

[27] Tipping, M.E. and Bishop, C.M. (1999) Probabilistic principal component analysis. Journal of the Royal Statistical Society: Series B, 61(3), 611-622.

[28] Yao, F., Müller, H.G., and Wang, J.L. (2005) Functional data analysis for sparse longitudinal data. Journal of American Statistical Association, 100(470), 577-590.

[29] Yao, F., Müller, H.G., Clifford, A.J., Dueker, S.R., Follett, J., Lin, Y., Buchholz B.A., and Vogel, J.S. (2003) Shrinkage estimation for functional principal component scores with application to the population kinetics of plasma folate. Biometrics, 59(3), 676-685.

[30] Zhang, L., Marron, J.S., Shen, H., and Zhu, Z. (2007) Singular value decomposition and its visualization. Journal of Computational and Graphical Statistics, 16(4), 833-854.

[31] Zhong, M., Sharma, S., and Lingras, P. (2004) Genetically designed models for accurate imputation of missing traffic counts. Transportation Research Recode: Journal of the Transportation Research Board, 1879, 71-79.