Bin TANG, Master[1]
E-mail: 2952071180@qq.com
Yao HU, Bachelor[1,2]
(Corresponding author)
E-mail: yhu1@gzu.edu.cn
Huan CHEN, Master[1]
E-mail: 2054879979@qq.com
[1] School of Mathematics and Statistics, Guizhou University
South Section of Huaxi Road, Huaxi District,
Guiyang 550025, China
[2] Guizhou Provincial Key Laboratory of Public Big Data,
Guizhou University
South Section of Huaxi Road, Huaxi District,
Guiyang, China

Intelligent Transport Systems
Original Scientific Paper
Submitted: 9 Dec. 2021
Accepted: 24 Feb. 2022

A FUNCTIONAL DATA APPROACH TO OUTLIER DETECTION AND IMPUTATION FOR TRAFFIC DENSITY DATA ON URBAN ARTERIAL ROADS

ABSTRACT

In traffic monitoring data analysis, the magnitude of traffic density plays an important role in determining the level of traffic congestion. This study proposes a data imputation method for spatio-functional principal component analysis (s-FPCA) and unifies anomaly curve detection, outlier confirmation and imputation of traffic density at target intersections. Firstly, the detection of anomalous curves is performed based on the binary principal component scores obtained from the functional data analysis, followed by the determination of the presence of outliers through threshold method. Secondly, an improved method for missing traffic data estimation based on upstream and downstream is proposed. Finally, a numerical study of the actual traffic density data is carried out, and the accuracy of s-FPCA for imputation is improved by 8.28%, 8.91% and 7.48%, respectively, when comparing to functional principal component analysis (FPCA) with daily traffic density data missing rates of 5%, 10% and 20%, proving the superiority of the method. This method can also be applied to the detection of outliers in traffic flow, imputation and other longitudinal data analysis with periodic fluctuations.

KEYWORDS

functional data; functional principal component analysis; traffic density; outliers; s-FPCA.

1. INTRODUCTION

Current intelligent transportation system (ITS) technologies capture a lot of traffic monitoring data, which has led to an increase in the demand for data analysis. Besides, macroscopic traffic flow theory considers that traffic speed, traffic density, traffic flow and other traffic characteristics of vehicle groups are the main features of traffic stream. Among them, traffic density is one of the indicators to measure the level of road service [1]. In addition, data quality is an important issue in traffic data analysis. Although these observations are usually recorded automatically by geomagnetic detectors and check points, there is a risk of data corruption due to short-term software or hardware failures, maintenance operations and data storage problems, which can create interruptions or outliers in the data records and lead to various types of subsequent modelling and identification of potential structural obstacles. Consequently, outlier identification and data imputation must be carried out prior to statistical analysis to minimise the impact of outlier data on subsequent modelling analysis and parameter estimation.

Functional data analysis (FDA) is very effective with infinite dimensional data objects such as curves, shapes and images. FDA methods have been developed in depth over the last 20 years, and a number of articles and books provide a comprehensive overview of FDA [2–3]. Chiou [4] first applied FDA to traffic flow data analysis, considered daily traffic flow trajectory as a stochastic function of time, proposed a dynamic functional hybrid prediction method combining functional prediction and probabilistic function classification for traffic

flow prediction, and subsequently used functional principal component analysis (FPCA) for estimating missing values of traffic flow data. The same researcher further extended it for anomalous curve monitoring graphical tool [5], and proposed a unified algorithm combining probabilistic function clustering algorithm with FPCA to propose a hybrid prediction [6] for missing value estimation. These studies reveal that the application of functional data analysis to traffic flow data analysis is more suitable and has advantages. Mu [7] applied functional principal component analysis to spectral data, calculated the Oja depth for the first and second principal component scores and used the box-line plot criterion to identify outliers for the Oja depth so as to determine the anomaly function. Chen [8] extended the computational space of Tukey's half-space depth based on the Hilbert regeneration kernel, proposed a new function-based statistical depth function and explored the outlier diagnosis methods for function-based data, mainly including the methods based on function-based depth, based on principal component analysis and based on function-based radial anomaly degree.

In the outlier detection algorithm, Hyndman and Shang [9] divided the anomalous curves of functional data into amplitude anomalous curves and shape anomalous curves, where amplitude anomalous curves are curves that deviate far from the mean and shape anomalous curves have different types of fluctuations than other curves, and proposed a visualisation method for the detection of anomalous curves. In practice, however, outlier curves may exhibit a combination of these features. In addition, outliers can greatly affect statistical results, such as distorting statistical models and deviating results. Chiou et al. [5] used logic rules to check for outliers, which were not changes in traffic status caused by traffic accidents or traffic jams, but only outliers caused by missing data during data collection. And due to the large amount of data, they also did not do subsequent anomaly detection and imputation, but directly deleted the outliers. On the one hand, this study improves the logic rules by using the threshold method for the determination of outliers to identify the existence of outliers more precisely. On the other hand, when the amount of data is small, it is obviously unreasonable to delete the anomalous curves directly or replace them with the mean curve, so it is proposed that the values of the anomalous curves will be subsequently processed using the imputation method to restore the original characteristics of the data as far as possible. Mondal and Rehena [10] propose a technique based on statistical model which identifies the temporal outliers in road traffic. Z-score and linear regression model are two statistical models have been used in combination for detection of temporal outliers. The proposed technique can be used to detect unusual traffic incident or sensors failure. Pu et al. [11] focuses on the detection of non-recurrent traffic anomaly caused by unexpected or transient incidents, such as traffic accidents, celebrations and disasters. Compared to existing approaches, it considers the spatial and temporal propagation of traffic anomalies from one road to other neighbour roads by proposing an STLP-OD framework.

It is also worth mentioning that in the identification of outliers using the three fundamental parameters of traffic flow, Chen et al. [12, 13] used two of the three parameters of traffic flow for curve fitting to obtain the highest and lowest thresholds of the basic graph curve, and anything outside the interval was judged to be an outlier, thus detecting the existence of traffic fault data. In this study we carry out the empirical plots for the fundamental diagram to identify non-anomalous areas based on the characteristics of time and traffic density, followed by determination of the outliers and imputation.

A comprehensive overview of the missing data problem is given by Schafer et al. [14]. A variety of imputation techniques have been developed over the last few decades [15–18], and there is an extensive literature discussing methods for inserting missing values for multivariate data [19] and longitudinal data [20–23]. There are specialised imputation methods for traffic data, including Kalman filtering methods, time series modelling [24], historical proximity estimation, lane distribution methods, spline regression estimation methods [25] and genetic design modelling [26]. More recently, Li and Chiou [6] applied the FPCA method to verify its superiority over probabilistic principal component analysis and Bayesian principal component analysis. Although historical estimation, nearest neighbour estimation and local regression estimation are common methods for missing value estimation, they all have certain drawbacks. They ignore the fact that traffic flows may fluctuate significantly from one day to the next and include random variations within the same day. Historical estimation uses global information in the historical data

closely related or near-neighbourhood information, while sample imputation uses local information intra-day traffic data for imputation. Although there have been many variational analysis methods to deal with missing values, to the best of our knowledge, the use of functional data analysis methods has not yet discussed how to apply to spatial attributes for longitudinal functional data for outlier identification and imputation of missing values.

In this study, anomalous curves are detected by highest density region (HDR) method, traffic density curves without anomalous values are clustered to confirm the potential structure corresponding to the anomalous curves, and the outliers on anomalous curves are detected by potential structure for the anomalous curves. The upstream and downstream density data and FPCA methods are then applied to imputation outliers, where each observed daily traffic flow curve is treated as a realisation of a random function subject to measurement error, until the anomalous curves have no outliers present.

This study proposes an outlier detection and imputation method for traffic density data that combines FDA to explore potential patterns through clustering and to fill in missing data in conjunction with spatial correlation of traffic data. K-means based clustering method effectively identifies classes with different contours and patterns of random variation in traffic flow trajectories, which helps to estimate missing values under different traffic flow patterns. The spatio-functional principal component analysis (s-FPCA) method uses spatial attributes of longitudinal functional data for outlier identification and imputation of missing values based on FPCA.

The article is organised as follows. Section 2 discusses the data sources and preliminary exploratory data analysis. In Section 3, FPCA and anomalous curve detection are described, as well as the unified algorithm for outliers detection and imputation. In Section 4, the algorithm is applied based on the dataset. In Section 5, the accuracy of the new algorithm is verified by simulation analysis. Section 6 summarises this paper and presents the future work.

2. DATA DESCRIPTION

We analysed the traffic density data collected by the check point located at the intersection of Changling South Road and Yangguan Avenue in Guiyang City, Guizhou Province. The traffic density data were collected for each 5-min interval from 1 March to 23 March in 2021, and each day's data contained 288 observation intervals, which were only consider the traffic density data in the two directions which are from south to north and north to south.

2.1 Preliminary exploratory data analysis

We divide the traffic density data for these 23 days into two groups, the working-day group with 17 days and the weekend group including the remaining 6 days. *Figure 1a* shows the total traffic density curve from 1 March to 23 March in 2021, and *Figures 1b, 1c and 1e* show the traffic density curves and mean curves for the working-days and weekends groups, respectively. There is also a small increase at 20:30 on both weekday and weekend days, which is related to people's daily routine. However, we can see that two of the daily traffic density curves in *Figure 1d* have significant anomalies, which motivates us to detect the anomalous curves and fill outliers and to consider clustering issues.

2.2 Missing data and outliers in the traffic stream

Data quality is an important issue in traffic data analysis. Although these data are usually recorded automatically by the check point, there is potential for data corruption to occur due to short term software or hardware failures, maintenance operations etc. From the anomalous observation curve in *Figure 1d*, it can be seen that there are some anomalous values of traffic density with a time interval of 5 minutes. Intermittent data recording or the presence of outliers creates various obstacles back to subsequent modelling and identification of potential random mechanisms, so it is important to identify outliers and imputation before statistical analysis. As for the outliers in the observation curves, one curve in *Figure 1d* contains outliers concentrated at the end of the morning peak and during the evening peak, and the other curve contains outliers mainly at the morning peak and during the time when the morning peak is approaching.

3. METHODOLOGY

In order to analyse the stochastic characteristics of the traffic density curve, the functional data approach is used in this study. The daily traffic density curves are considered as implementations of random functions sampled from random processes, where each potential process represents a subgroup with different characteristics. Functional data

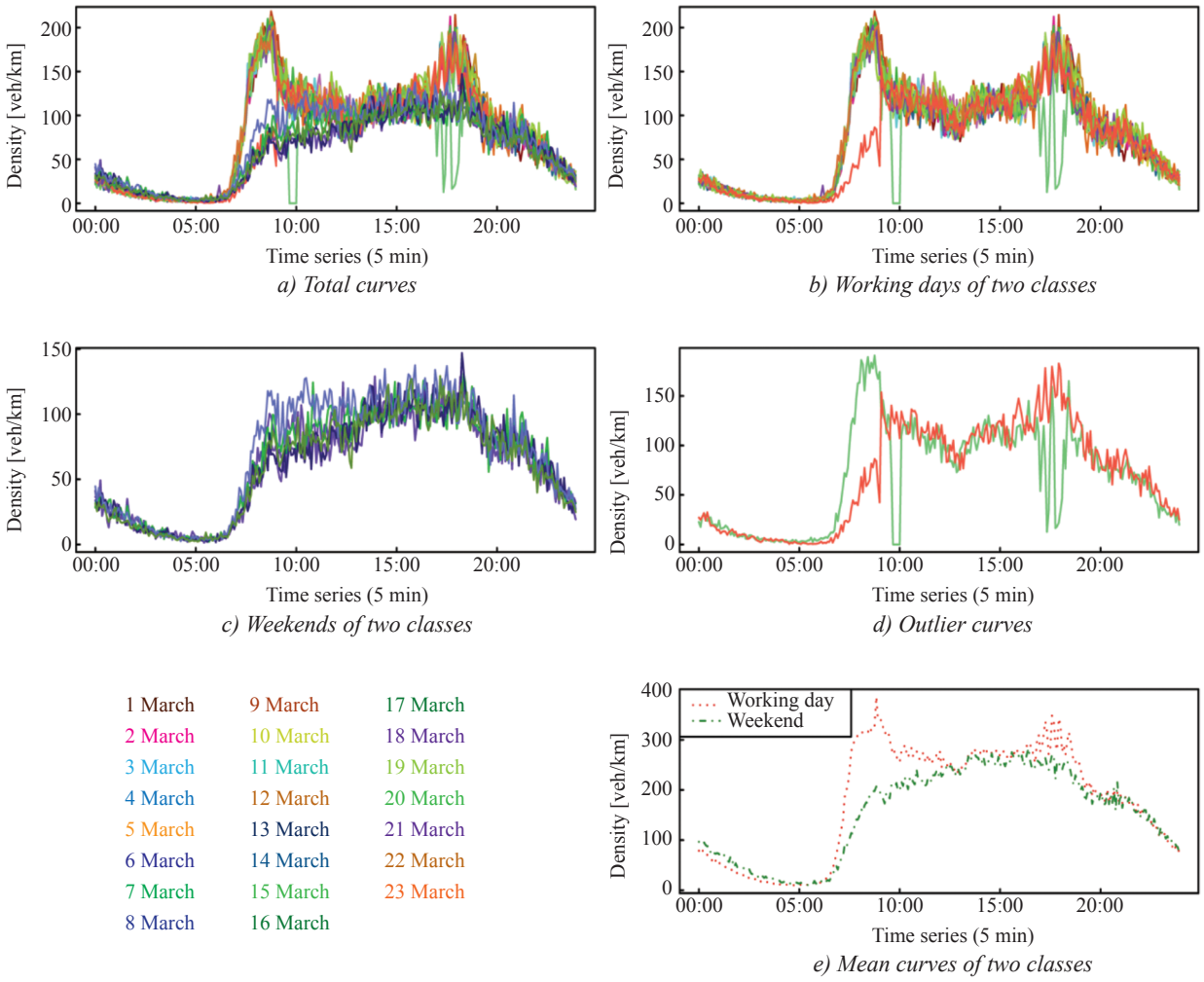


Figure 1 – Traffic density curves at the intersection of Guiyang city

analysis is first performed to detect anomalous curves as well as to identify the corresponding outliers, and then to fill in the outliers by using a clustering algorithm as well as s-FPCA, which constitutes a unified algorithm for anomaly detection and filling. Anomalous curve detection and identification of outliers is first performed, followed by imputation of outliers by using a clustering algorithm as well as s-FPCA, which constitutes a unified algorithm for outliers detection and imputation.

3.1 Functional principal component analysis

Consider a random function Y in $L^2(H)$ for the daily traffic density trajectory, where $L^2(H)$ denotes a Hilbert space of squared integrable functions on the closed time interval H=[0,T]. The inner product of two functions d and l in $L^2(H)$ is defined as $\langle d,l\rangle = \int_H d(t)l(t)dt$ with the norm $\|\cdot\| = \langle\cdot,\cdot\rangle^{1/2}$. Suppose Y has a mean function $\mu(t)=E\{Y(t)\}$, and $\Gamma(s,t)=\mathrm{cov}\{Y(s),Y(t)\}=E\{(Y(t)-\mu(t))(Y(s)-\mu(s))\}$ is a covariance function. The covariance function can be represented by the eigen decomposition as $\Gamma(s,t)=\Sigma_r\lambda_r\varphi_r(s)\varphi_r(t)$, where the non-negative eigenvalues are $\lambda_1\geq\lambda_2\geq...\geq 0$ and satisfy $\Sigma_r\lambda_r<\infty$, the standard orthonormal eigenfunctions $\{\varphi_r\}$ corresponding to the eigenvalues $\{\lambda_r\}$. The random traffic density trajectory Y can be expressed by the *K-L* expansion as:

$$Y(t)=\mu(t)+\sum_{t=1}^{\infty}\xi_r\varphi_r(t) \tag{1}$$

where the random effects $\xi_r=\langle X-\mu,\varphi_r\rangle$ are uncorrelated function principal component scores (FPC scores) with zero mean and variances λ_r. Furthermore, for the i^{th} traffic density trajectory $Z_i(t)$ with measurement error, the observation at time point T_{ij} can be expressed as:

$$Z_{ij}=Y_i(T_{ij})+\varepsilon_{ij}=\mu(t)+\sum_{r=1}^{\infty}\xi_{ir}\phi_r(T_{ij})+\varepsilon_{ij} \tag{2}$$

where random error ε_{ij} with zero mean and variances $Var\{\varepsilon_{ij}\}=\sigma^2$, and uncorrelated with $\xi_r=\langle Y-\mu,\varphi_r\rangle$.

To estimate the nonparametric smooth mean function $\mu(t)$ and covariance function $\Gamma(s,t)$ in *Equation 1*, we apply the one-dimensional linear smoothing to the observed data $\{(T_{ij},Z_{ij}),j=1,...,m_i,i=1,...,n\}$ to estimate $\mu(t)$. Then the covariance function Γ is estimated by applying two-dimensional scatterplot smoother to the raw covariances $\Gamma(s,t)$. The estimate of the characteristic function $\{\varphi_r\}$ is obtained by $\hat{\Gamma}(s,t)=\Sigma_r\hat{\lambda}_r\hat{\varphi}_r(s)\hat{\varphi}_r(t)$, and when the discrete observations of the function data are densely sampled, the random effect ξ_{ir} can be estimated by numerical integration $\xi_{ir} = \int (Z_{ij} - \mu(t))\varphi_r(t)dt$.

3.2 Abnormal curve detection of functional data

HDR was first proposed by Hyndman and is analysed and plotted from the first two FPC scores of functional data. The 2D HDR boxplot is constructed from the 2D kernel density estimate $\hat{f}(z)$, and the HDR is defined as follows.

$$R_\alpha = \{z:\hat{f}(z) \geq f_\alpha\} \tag{3}$$

where $\int_{R\alpha} \hat{f}(z)dz = 1 - \alpha$, and f_α is the region with probability 1-α. Points in this region have a higher density than points outside this region, hence the term 'highest density region'. The 2D HDR boxplot has a mode, as well as 50% of the inner region and 99% of the outer region, with the mode being the highest density points and the points outside the outer region being considered outliers.

The binary principal component HDR boxplot is mapped to a function space to obtain a functional bagplot, where the mode corresponds to the reference curve that has the highest density, and the inner and outer regions are the regions that contain the curves corresponding to the points in them, respectively. Taking the first two FPC scores as the horizontal axis and the vertical axis respectively, the binary principal component HDR boxplot is obtained. The values deviating from most points correspond to the abnormal curves detected by the functional bagplot.

3.3 Unified algorithm for outliers detection and imputation

In previous papers on detecting traffic flow anomalies, the original curve was deleted after the presence of the anomalous curve was detected when the amount of data was large. However, in the case of a small amount of data, it is necessary to restore the original data as much as possible to reduce the impact on the subsequent analysis. In this study a unified algorithm for outliers detection and imputation is proposed, and the algorithm steps are as follows.

Step 1: Anomalous curves detection and different structures division

After judging the anomalies according to the empirical plots for the fundamental diagram of traffic flow and excluding the traffic fault situation, the original data are tested for anomalous curves. In classification based on clustering features for k-means clustering of non-anomalous curves, density structure of different classes *c* can be expressed as:

$$Z^{(c)}(t) = \mu^{(c)}(t) + \sum_{t=1}^{M} \xi_r^{(c)}\varphi_r^{(c)}(t) \tag{4}$$

Step 2: Outliers detection

To identify outliers according to the structures classified in step 1, first we need to confirm the non-anomalous area when we perform the outliers determination. The threshold detection method commonly used in traffic flow data anomaly detection, and the threshold value of traffic density at the first time point of the day is confirmed by the historical maximum and minimum values as shown in *Equation 5* below.

$$Z_{i1}^{max} = \max(Z_{i1}) + \sigma_1;\ \ Z_{i1}^{min} = \min(Z_{i1}) + \sigma_1 \tag{5}$$

where σ_1 is the standard deviation value at the first point in the non-anomalous curves. The enclosed interval by $(Z_{i1}^{max},Z_{i2}^{max},Z_{ij}^{max},...,Z_{in}^{max})$ and $(Z_{i1}^{min},Z_{i2}^{min},Z_{ij}^{min},...,Z_{in}^{min})$ is the obtained non-anomaly region according to the historical maximum and minimum threshold. Then, the points outside this region are considered as anomalies.

Step 3: Outliers imputation with s-FPCA

Traffic flow data has its own characteristics. Firstly, temporal correlation shows that traffic flow parameters of road sections fluctuate with time following a certain trend, and dynamic traffic flow changes continuously with time and there is a certain trend of fluctuations. Secondly, historical correlation shows that traffic flow parameters present similar characteristics at the same time of different days, and the traffic flow cycle change pattern is more obvious. In this study we consider clustering algorithms that divide traffic density curves into weekdays and weekends, reflecting the

existence of different potential structures on different days. Temporal and historical correlations can be expressed through different potential structures as *Equation 4*. Missing values can be predicted and imputed by different structures for traffic flow parameters.

In addition, big data on traffic flow is characterised by spatial correlation, as the traffic flow parameters of a road segment are influenced by the upstream and downstream segments, which can be expressed in terms of correlation as $\tilde{Z}_{ij} = a_1 Z_{ij}^1 + a_2 Z_{ij}^2 + b$, where the $\tilde{Z}_{ij}$ is the traffic density at the j^{th} 5-minute interval on day i that the target intersection needs to be imputed. Z^1 and Z^2 are the traffic density values upstream and downstream of the target intersection, respectively, and a_1, a_2 and b are the linear coefficients fitted to the spatial correlations, respectively. Due to the spatial correlation of the traffic flow big data, there is a high accuracy rate for traffic flow parameter prediction. However, due to the strong volatility of the traffic flow, for some points the fluctuations will have a greater impact on the imputation results. In addition, in the case of the existence of anomalous missing upstream and downstream data, we can consider the following method for the imputation of the correlation. Based on the consideration of the potential process, the correlation formula is expressed as:

$$\begin{aligned}\tilde{Z}^{(c)}(t) = a_1 &\left\{\mu_1^{(c)}(t) + \sum_{t=1}^{M} \xi_{1r}^{(c)} \varphi_{1r}^{(c)}(t)\right\} \\ &+ a_2 \left\{\mu_2^{(c)}(t) + \sum_{t=1}^{M} \xi_{2r}^{(c)} \varphi_{2r}^{(c)}(t)\right\} + b\end{aligned} \tag{6}$$

where $\left\{\mu_1^{(c)}(t) + \sum_{t=1}^{M} \xi_{1r}^{(c)} \varphi_{1r}^{(c)}(t)\right\}$ is the upstream FPCA with structure c, downstream FPCA with structure c can be expressed as $\left\{\mu_2^{(c)}(t) + \sum_{t=1}^{M} \xi_{2r}^{(c)} \varphi_{2r}^{(c)}(t)\right\}$. a_1, a_2 and b are the linear coefficients fitted by correlations, respectively. The s-FPCA imputation method is based on imputation with correlations and iterating through the FPCA until there are no outliers in the anomalous curve.

Step 4: Iteration until convergence

Repeat steps 1 through 3 until the results converge to the initial anomalous curve being non-anomalous.

4. EXPERIMENT AND DISCUSSION

Outliers in the velocity, flow and density data were not found through the empirical plots for the fundamental diagram of traffic flow shown in *Figure 2*, where the different points indicate parameter values at different time intervals on different dates. The Q-K diagram shows the free-flowing part of the traffic flow and does not present any outliers with a large degree of dispersion. In addition, the V-K and V-Q diagrams show discrete speed values at low density and flow values. In practice, in the early morning hours when traffic flow is low, outlier speed values often occur. In contrast, when traffic flow and traffic density are greater, vehicle speed is more stable. In addition, since our observed data objects are located on urban arterial roads, the traffic condition is good as seen by the Q-K plot, and subsequent outliers processing is required.

4.1 Anomalous detection and non-anomalous curve clustering

In *Figure 3*, the grey area in the right panel represents 99% of the non-abnormal data area, and traffic density curves for 11 March and 23 March were detected as anomalies. The traffic density on 11 March was a significant data anomaly, with outliers present during both mini-peaks, at 9:10 a.m. and 2:30 p.m., respectively. Outliers were also present during the early morning to morning rush hours on 23 March, with a spike in traffic density occurring around 9:00 am. The functional bagplot on the left panel is a mapping of the binary principal component boxplot on the functional space, and the grey area on the right panel corresponds to the grey function area space in the left panel of *Figure 3*. The points with asterisks in the right figure correspond to the functional mean curve, and the two outliers in the right panel correspond to the traffic density curve on 11 March and 23 March in the functional curve in the left panel of *Figure 3*.

The k-means curve clustering was performed on the remaining 21 days of data, and the results were category 1 for weekdays, totalling 15 days, and category 2 for weekend days, totalling 6 days. The 2 anomalous curves are all working days, then the potential structure of weekdays is applied in the imputation method considering clustering. Next, this study performs the identification of non-anomalous regions by confirming thresholds, followed by the identification of outliers in anomalous curves.

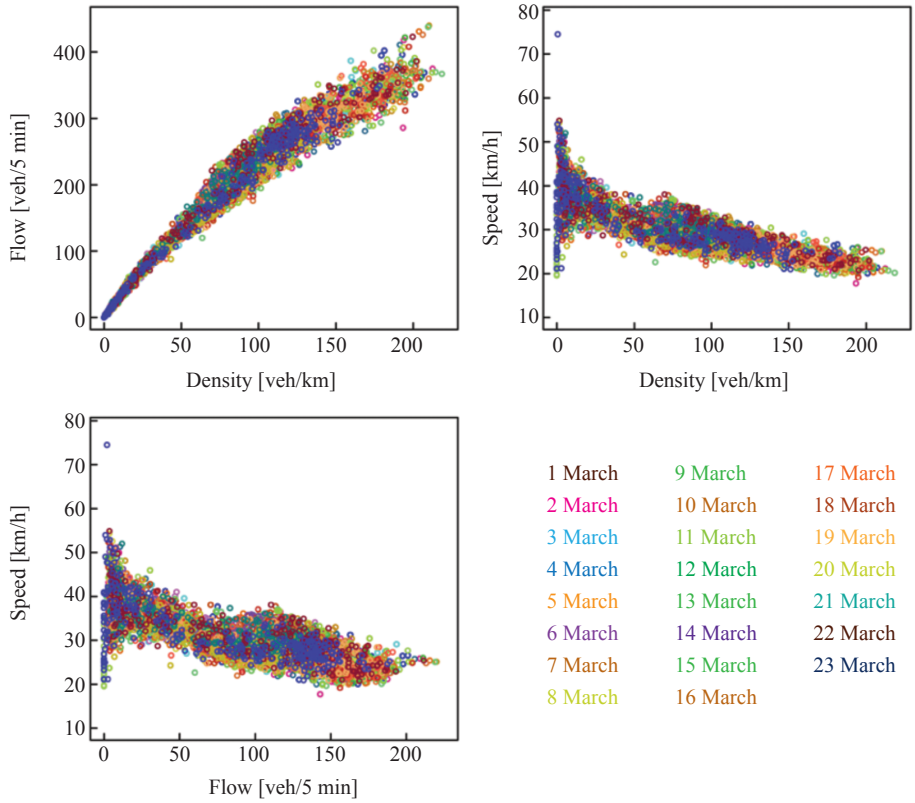


Figure 2 – Empirical plots for the fundamental diagram of traffic flow

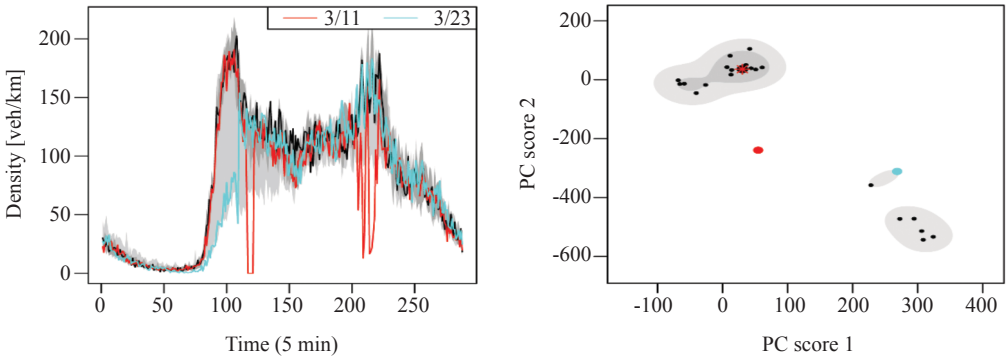
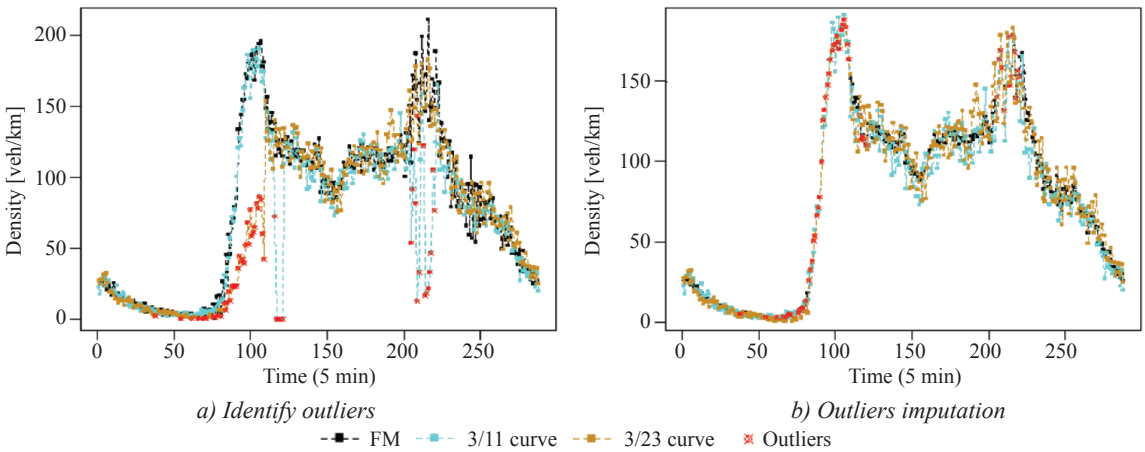


Figure 3 – Detection of anomalous curves



a) Identify outliers *b) Outliers imputation*

Figure 4 – Identification and imputation of the outliers

4.2 Confirmation of outliers and imputation in the anomalous curve

In previous studies on anomaly curve detection, only the confirmation of anomaly curves was done without subsequent anomaly detection and imputation. For this research observation data object, when the amount of data is small, it is obviously unreasonable to delete the anomaly curve directly or replace it with the mean curve, so the detection of anomalous points in the anomalous curve and the subsequent imputation are very important. As shown in the left panel of *Figure 4*, the final confirmed outliers can be seen to be relatively complete, and the subsequent imputation work is required.

4.3 Outliers imputation with s-FPCA

Chiou et al. [5] classify missing patterns of traffic flow data into point missing (PM), interval missing (IM) and mixed missing (PM/IM). Missing points in the PM set are completely isolated and randomly dispersed, while the IM set contains randomly distributed unobserved intervals. Mixed PM/IM is a combination of PM and IM, where we obtain observations for the IM interval missing pattern only. The right panel of *Figure 4* shows the imputation results, which can be seen to capture the trend of the traffic density. Then s-FPCA should be evaluated by simulation analysis.

5. SIMULATION WITH S-FPCA

This section verifies the effectiveness of the algorithm by applying the actual traffic data at the intersection and its upstream and downstream.

5.1 Dataset of the simulation

For our observation of missing values of traffic data, not all data intervals have missing values. Here we still use the traffic density data at the intersection of South Changling Road and Yangguan Avenue and the traffic density of two directions upstream and downstream of this intersection for the simulation analysis. The simulation analysis was performed by randomly screening the 5 March 2021 data for 5%, 10% and 20% missing data points, and the simulated missing pattern was PM/IM missing, which is closer to the real situation. The missing pattern is shown in *Figure 5*, where the hollow points are randomly set missing points. We use RMSE, MAPE and MAE as criteria for judging the different imputation methods.

$$RMSE = \sqrt{\frac{1}{m}\sum_{j=1}^{m} |\hat{y}_i(t_{ij}) - y_i(t_{ij})|^2}$$
$$MAPE = \frac{1}{m}\sum_{j=1}^{m} \left|\frac{\hat{y}_i(t_{ij}) - y_i(t_{ij})}{y_i(t_{ij})}\right| \quad (7)$$
$$MAE = \frac{1}{m}\sum_{j=1}^{m} |\hat{y}_i(t_{ij}) - y_i(t_{ij})|$$

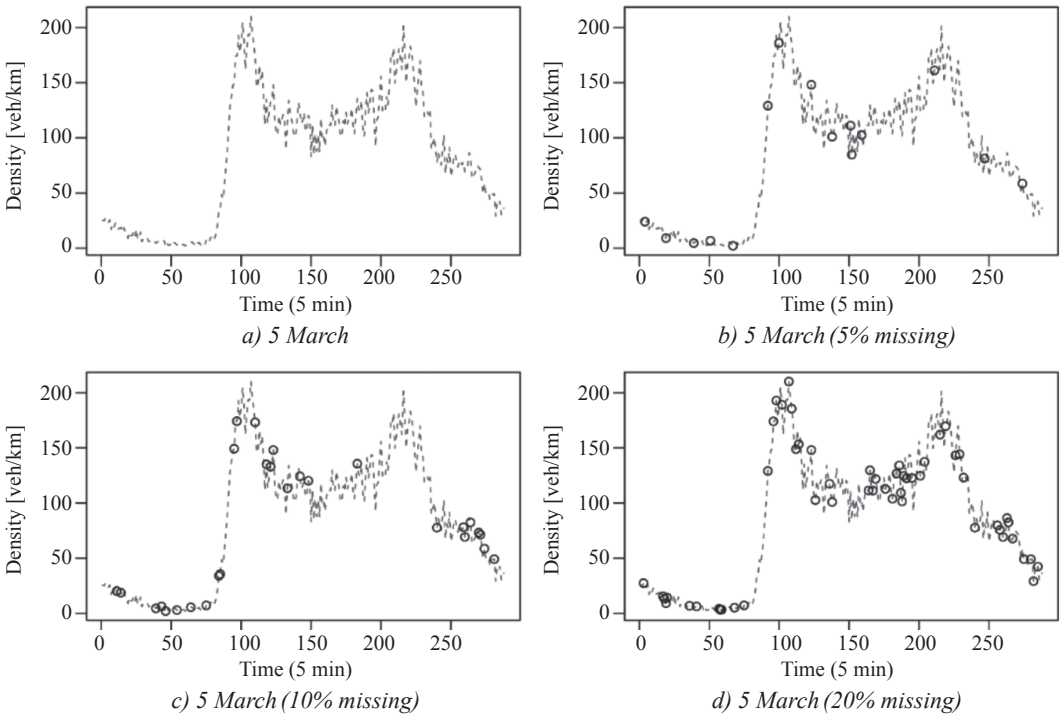


Figure 5 – Visualisation of the lack of traffic density on 5 March 2021

Table 1 – Missing imputation simulation results

Missing Ratio	Classification	Method	RMSE	MAPE	MAE
5%	None	FM	23.33850	0.30316	17.06415
5%	2-clusters	FM	18.81112	0.27195	14.07751
5%	2-clusters	FPCA	10.91388	0.25904	8.41512
5%	2-clusters	Spatial	19.97627	0.83074	16.70494
5%	2-clusters	s-FPCA	10.01067	0.30474	7.70696
10%	None	FM	23.71158	0.26451	18.07770
10%	2-clusters	FM	20.01900	0.24986	15.76949
10%	2-clusters	FPCA	12.27638	0.18529	9.14653
10%	2-clusters	Spatial	14.76045	0.62741	11.81379
10%	2-clusters	s-FPCA	11.18310	0.18613	8.00389
20%	None	FM	25.68154	0.19214	19.63514
20%	2-clusters	FM	21.45012	0.19052	17.23770
20%	2-clusters	FPCA	12.11154	0.12064	9.08126
20%	2-clusters	Spatial	19.25250	0.40251	15.41977
20%	2-clusters	s-FPCA	11.20615	0.11630	8.27620

where $y_i(t_{ij})$ are the actual observation data, $\hat{y}_i(t_{ij})$ are the expected observations from our calculations, i is the simulation on a fixed day, in this case on March 5, and m is the total number of missing points.

From *Table 1* we can see the results of the final simulation analysis for missing value imputation. Firstly, for the missing ratios of 5%, 10% and 20%, the results of imputation without clustering differed greatly from the original observations, as shown by the fact that the RMSE and MAE were larger without clustering than after clustering with the same FM imputation method to fill the missing values, and MAPE did not have a great advantage. Due to the influence of spatial correlation of traffic data, the results of filling by upstream and downstream density data are much superior compared to FM. Since linear correlation tends to focus more on larger traffic density values, MAPE has larger interpolation results with spatial correlation only. However, RMSE is consistent with MAE performance, so the imputation effect of spatial correlation is not the most ideal. We can see from MAPE that the results of prediction by upstream and downstream spatial correlation only have great uncertainty when the value of traffic flow parameter is small, which is also why the process of fitting linear relationships is more concerned with the fit of large parameter values, so the result of combining spatial correlation with FPCA is the closest to the original value.

Finally, it can be observed from *Figure 6* that the imputation of missing values by s-FPCA has the smallest RMSE and MAE regardless of 5%, 10%

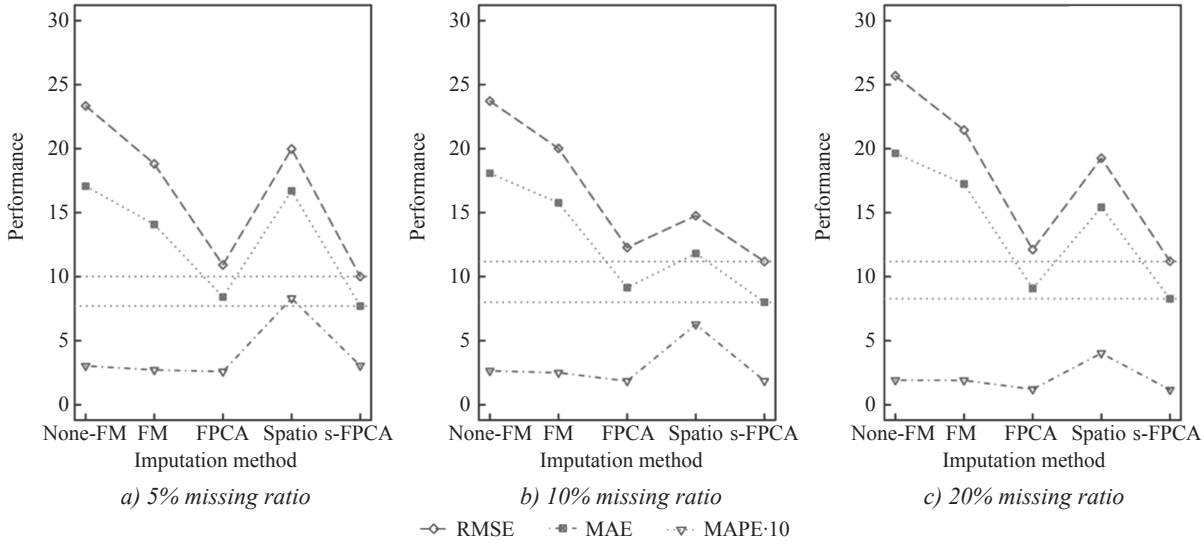


Figure 6 – Simulation analysis and visualisation of missing data imputation

or 20% missing ratio, and the MAPE also has the smallest value at 20% missing ratio. The results indicate that the initial imputation by spatial relations followed by FPCA iteration gives better results and better captures the fluctuations of traffic flow curves.

6. CONCLUSION

This study proposes an outlier detection and imputation method for traffic density data that combines FDA to explore potential patterns through clustering and to fill in missing data in conjunction with spatial correlation of traffic data. FPCA can consider important functional data features, including mean functions and random variations of single curves. The method is nonparametric, without any distributional assumptions and can adaptively fit the mean and covariance functions of individual clusters. K-means based clustering method effectively identifies classes with different contours and patterns of random variation in traffic flow trajectories, which helps to estimate missing values under different traffic flow patterns. Simulation studies show that considering traffic flow classification and spatial characteristic are beneficial to improve missing value imputation accuracy. This study proposes a method for anomaly detection of traffic data, in which anomaly curves are judged with HDR, and then outliers are judged by the historical maximum-minimum threshold. The simulation analysis results indicate that s-FPCA has better imputation results and better captures the fluctuations of traffic flow curves than the initial imputation by spatial relations followed by iteration by FPCA, which is also applied in the detection and imputation of outliers in other traffic parameters and longitudinal data.

Deep learning-based traffic flow prediction and imputation is an emerging method for traffic flow imputation, and various deep neural networks have been developed for the imputation of missing traffic flow data, such as self-attention mechanism, graph self-encoder [27] and tensor decomposition theory [28]. In particular, Li et al. [29] propose a model that involves the use of two parallel stacked autoencoders that can simultaneously consider the spatial and temporal dependencies.

Current research shows that prediction by neural networks can reduce the impact of missing or erroneous data by implementing an interpolation strategy, and in typical traffic studies, deep neural network methods use road network information for prediction and imputation. In contrast, our functional data approach considers the underlying patterns of daily traffic flow trajectories at individual vehicle inspection stations and provides an imputation method using FDA that is conceptually simple and relatively easy to compute. In the future, it will be a worthwhile research pursuit to combine the method with neural networks to further improve prediction performance.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (12161016, 11661018) and Guizhou Data Drive Modelling Learning and Optimisation Innovation Team ([2020]5016). The experimental data were provided by the Guiyang Public Security Traffic Management Bureau of Guizhou Province.

唐斌，硕士[1]
电子邮箱：2952071180@qq.com
胡尧，本科[1,2]
（通讯作者）
电子邮箱：yhu1@gzu.edu.cn
陈欢，硕士[1]
电子邮箱：2054879979@qq.com
[1] 贵州大学数学与统计学院
中国贵州省贵阳市花溪区花溪大道南段
[2] 贵州大学贵州省公共大数据重点实验室
中国贵州省贵阳市花溪区花溪大道南段

基于函数型数据分析的城市主干道交通密度异常值检测与插补方法

摘要：

在交通监测数据分析中，交通密度的大小对确定交通拥挤程度起着重要作用。本研究提出了一种*s-FPCA*的数据插补方法，将异常曲线检测、异常值确认和目标交叉口交通密度插补相结合。首先，根据从函数型数据分析中获得的二元主成分得分进行异常曲线检测，然后通过阈值法判定异常值的存在。其次，提出了一种基于上下游的改进的缺失数据插补方法。最后，对实际交通流密度数据进行模拟研究，在日交通密度数据缺失率分别为*5%*、*10%*和*20%*的情况下，与*FPCA*相比，*s-FPCA*的插补精度分别提高了*8.28%*、*8.91%*和*7.48%*，证明了该方法的优越性。另外该方法还可用于交通流量异常值的检测、插补和其他周期性波动的纵向数据分析。

关键词：

函数型数据；函数型主成分分析；交通密度；异常值；*s-FPCA*

REFERENCES

[1] Arasan VT, Dhivya G. Methodology for Determination of Concentration of Heterogeneous Traffic.

Journal of Transportation Systems Engineering and Information Technology. 2010;10(4). doi: 10.1016/S1570-6672(09)60052-0.

[2] Ramsay JO, Silverman BW. *Functional Data Analysis*. New York: Springer; 2005.

[3] Wang J-L, et al. Functional data analysis. *Annual Review of Statistics and Its Application.* 2016;3: 257-295. doi: 10.1146/annurev-statistics-041715-033624.

[4] Chiou J-M. Dynamical functional prediction and classification, with application to traffic flow prediction. *The Annals of Applied Statistics*. 2012;6(4). doi: 10.1214/12-AOAS595.

[5] Chiou J-M, et al. A functional data approach to missing value imputation and outlier detection for traffic flow data. *Transportmetrica B: Transport Dynamics.* 2014;2(2). doi: 10.1080/21680566.2014.892847.

[6] Li PL, Chiou J-M. Functional clustering and missing value imputation of traffic flow trajectories. *Transportmetrica B: Transport Dynamics.* 2020;9(1). doi:10.1080/21680566.2020.1781706.

[7] Mu W. Application of functional data anomaly detection in spectral data. *Xiamen University*, 2019.

[8] Chen J. Improvement and application of abnormal value diagnosis method of functional data. *Jiangxi University of Finance and Economics*. 2020. doi: 10.27175/d.cnki.gjxcu. 2020.000382.

[9] Hyndman RJ, Shang HL. Rainbow plots, bagplots, and boxplots for functional data. *Journal of Computational and Graphical Statistics*. 2010;19(1). doi: 10.1198/jcgs.2009.08158.

[10] Mondal MA, Rehena Z. Road traffic outlier detection technique based on linear regression. *Procedia Computer Science.* 2020;171(C): 2537-2555. doi: 10.1016/j.procs.2020.04.276.

[11] Pu J, et al. STLP-OD: Spatial and temporal label propagation for traffic outlier detection. *IEEE Access*. 2019;(7): 63036-63044. doi: 10.1109/ACCESS.2019.2916853.

[12] Chen K, Zou Q. [A traffic flow anomaly mining method incorporating time-correlated factor curve fitting]. *Computer Engineering and Design*. 2013;34(07): 2561-2565. Chinese.

[13] Lu M-W, et al. [*A traffic data pre-processing method based on curve-fitting anomaly detection*]. Database Professional Committee of China Computer Society; 2006. p. 642-646. Chinese.

[14] Schafer JL, Graham JW. Missing data: Our view of the state of the art. *Psychological Methods*. 2002;7(2): 147-177. doi: 10.1037/1082-989X.7.2.147.

[15] Collins LM, Schafer JL, Kam CM. A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*. 2001;6(4): 330-351. doi: 10.1037/1082-989X.6.4.330.

[16] Rubin DB. Multiple imputation for nonresponse in surveys. *John Wiley & Sons*; 2004. p. 81.

[17] Booth DE. Analysis of incomplete multivariate data. *Technometrics*. 2000;42(2): 213-214 doi: 10.1080/00401706.2000.10486013.

[18] Schlittgen R. Analysis of incomplete multivariate data. *Computational Statistics and Data Analysis.* 1999;30(4): 478-479. doi: 10.1016/S0167-9473(99)90025-7.

[19] Beale EML, Little RJA. Missing values in multivariate analysis. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1975;37(1): 129-145. doi: 10.1111/j.2517-6161.1975.tb01037.x.

[20] Laird NM. Missing data in longitudinal studies. *Statistics in Medicine*. 1988;7(1-2): 305-315. doi: 10.1002/sim.4780070131.

[21] Little RJA. Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association.* 1995;90(431): 1112-1121.

[22] Little RJA, Rubin DB. Statistical analysis with missing data. *John Wiley & Sons, Inc;* 2002. doi: 10.1002/9781119013563.

[23] Molenberghs G. Applied longitudinal analysis. *Journal of the American Statistical Association.* 2005;100(470). doi: 10.1198/jasa.2005.s24.

[24] Nihan NL. Aid to determining freeway metering rates and detecting loop errors. *Journal of Transportation Engineering.* 1997;123(6): 454-458. doi: 10.1061/(ASCE)0733-947X(1997)123:6(454).

[25] Chen C, et al. Detecting errors and imputing missing data for single-loop surveillance systems. *Transportation Research Record*. 2003;1855(1): 160-167. doi: 10.3141/1855-20.

[26] Zhong M, Sharma S, Lingras P. Genetically designed models for accurate imputation of missing traffic counts. *Transportation Research Record*. 2004;1879(1): doi: 10.3141/1879-09.

[27] Zhang W-B, et al. [Traffic flow data restoration model for road networks based on self-attention mechanism and graph self-encoder]. *Transportation Systems Engineering and Information*. 2021;21(04): 90-98. doi: 10.16097/j.cnki.1009-6744.2021.04.011. Chinese.

[28] Lu W-Q, et al. [Lane level traffic flow data restoration algorithm based on tensor decomposition theory]. *Journal of Jilin University (Engineering and Technology Edition)*. 2021;51(05): 1708-1715. doi: 10.13229/j.cnki.jdxbgxb20200535. Chinese.

[29] Li L, et al. Estimation of missing values in heterogeneous traffic data: Application of multimodal deep learning model. *Knowledge-Based Systems*. 2020;194: 105592. doi: 10.1016/j.knosys.2020.105592.