ACADEMIC
World Academic Union

ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 14, No. 3, 2019, *pp*.184-194

Functional clustering with application to air quality analysis

Ming He, Hairong Li, Xiaoxin Zhu, Chunzheng Cao[1]

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China

(Received March 21 2019, accepted June 20 2019)

Abstract. Based on the air quality status of 161 cities in China, this paper studies the temporal and spatial distribution characteristics of PM2.5 concentration of major pollutants affecting air quality index (AQI). We use improved functional clustering analysis methods and add priori information about location and human factors to make the clustering results more accurate. The improved functional clustering model is compared with the basic sparse data function clustering method, k-centres functional clustering method, functional principal component analysis and traditional K-means clustering method by repeated simulation. Finally, we use the PM2.5 concentration of selected 161 cities in China as an illustrative example.

Keywords: air quality index, PM2.5 concentration, functional clustering, priori information

1. Introduction

Air quality influences human health and economic development. Nowadays air quality is measured by air quality index (AQI), which is typically a temporal-spatial data. This research is motivated by an air quality influences human health and economic development.

Many existing studies have analyzed air quality but they are limited to use simple statistical methods and spatial correlations [1, 2]. And some researches just consider dozens of cities such as Chen used EPLS method to analysis air quality of 31 provincial capitals in China mainland [3]. Hamedian [4] mainly used fuzzy C-mean clustering to find the main pollutants which can influence the air quality. In this article, air quality is measured by Air Quality Index (AQI) and PM2.5, PM10, SO2, O3, CO, NO2 six pollutants of 161 cities in China. We leverage new statistical methods for estimating and describing air quality trends and distribution that can be used to inform about spatial and temporal distribution characteristics.

Cluster analysis is the art of identifying groups in data. Traditional clustering methods are focus on multivariate data and many clustering algorithms have been proposed when the data are curves or functions. In this context, Functional Data Analysis has received increasing attention recently [5].

Several clustering methods for functional data have been researched in recent years. The two-stage approach was proposed by Maharaj [6] and used by Iorio et al. [7] to handle time course data with observed measurements. P-splines smoothers was used to model the observed measurements and then to cluster functions by the optimal spline coefficients. They added penalty term based on the general basis expansion and fitted the curves well by choosing smoothing parameter. Traditional approaches based on clustering basis coefficients choose the same basis functions for all clusters to use the fitted coefficients to be clustered. There are some problems because basis functions should be chosen then the fitted coefficients can adequately adapt cluster differences. For the model-based clustering method, Same´ and Bouveyron used this approach in mixture model based on high dimensional data [8, 9]. Basically, the model parameters are always estimated by the maximum likelihood method solved by an Expectation-Maximization (EM) algorithm [10]. When the observations are sparse, irregularly spaced, or occur at different time points for each subject, James and Sugar proposed a particularly effective model-based approach for clustering functional data [11]. They produced low-dimension representation of the curves and then provided low-dimension graphical representations to show some direct clustering results in the pictures. In fact, various model-based approaches are under certain probability model assumptions. Just considering the information of curves themselves without some correlated variables may not cluster well. Chiou and Li [12] proposed a k-centres functional clustering method which can greatly improve cluster quality compared with the conventional clustering algorithms. The k-centres functional clustering method does not rely on any distribution assumptions and the mean and covariation structures of each cluster are explored using this approach.

[1] Corresponding author. *E-mail address*: caochunzheng@163.com.

This study is concemed with functional data clustering where the number of observations is 161 cities in China and the recording times are the same for individuals. Thus, the algorithm for sparse samples should be improved and we consider to add the position information as prior when fit the air quality curves. A logistic function and its similar form were considered to be probability by many model-based functional clustering approaches [13]. It is worth noting that the determination of a clustering technique is even more difficult under the possible presence of outlying curves. One possibility to improve the robustness of clustering algorithm is through the application of trimming tools. In this study, low-dimension representation and visual exhibition are considered. Combining logistic prior information and robust trimming tools, we compare with two typical functional clustering methods: k-centres functional clustering and sparsely functional clustering.

This paper is organized as follows. In section 2, we define the proposed model, also detail the method of parameter estimations, the curve clustering and model selection. In section 3, we compare the improved functional clustering model with other clustering methods through repeated simulations. In section 4, we conducted a cluster analysis of PM2.5 concentrations in selected 161 cities in China, and compared the differences in air quality between different types of cities. Finally, some conclusions are presented in section 5.

2. Methodology

2.1. The model

Let $f_i(t)$ be the value for the i-th smooth underlying curve. The observed data can be expressed as

$$y_{ij} = f_i(t_{ij}) + \varepsilon_{ij}, \quad i = 1,\ldots,n,\ j = 1,\ldots,n_i, \tag{1}$$

where $\boldsymbol{y}_i = (y_{i1},\ldots,y_{in_i})^T$ are the vectors of observed values at time points $\boldsymbol{t}_i = (t_{i1},\ldots,t_{in_i})^T$ and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1},\ldots,\varepsilon_{in_i})^T$ are measurement errors following $N(0,\sigma^2\boldsymbol{I}_{n_i})$. The subject-specific random function $f_i(t)$ is a Gaussian process, which can be approximated as

$$f_i(t) = \mathbf{s}(t)^T\boldsymbol{\eta}_i, \tag{2}$$

where $\boldsymbol{s}(t)$ is p-dimensional vector of spline basis function with $\boldsymbol{\eta}_i$ is coefficient vector of the spline basis, which can be modeled using the following Gaussian distribution

$$\boldsymbol{\eta}_i\,|_{z_{ik}=1} = \boldsymbol{\mu}_k + \boldsymbol{\gamma}_i, \quad \boldsymbol{\gamma}_i \sim N(\mathbf{0},\boldsymbol{\Gamma}_k), \tag{3}$$

where the latent label z_{ik} denotes the cluster membership vector for the i-th individual, when $z_{ik} = 1$, $f_i(t)$ belongs to the k-th cluster and $z_{ik} = 0$ otherwise.

In model-based clustering it is assumed that the observations $\boldsymbol{y}_1,\ldots,\boldsymbol{y}_n$ follow a mixture distribution with K components. In addition, $\boldsymbol{z}_i = (z_{i1},\ldots,z_{iK})^T$ follows a multinomial distribution with parameter $(\pi_{i1},\ldots,\pi_{iK})^T$ and π_{ik} is the probability of the i-th observation belongs to the k-th cluster. Suppose there exists a p_w dimensional covariates $\boldsymbol{w}_i = (1, w_{1i},\ldots,w_{p_w-1,i})^T$ which can influence the categorical latent variable $\boldsymbol{z}_i$ through a logistic model

$$\pi_{ik} \triangleq P(z_{ik}=1) = \frac{\exp(\mathbf{w}_i^T\mathbf{v}_k)}{1+\sum_{j=1}^{K-1}\exp(\mathbf{w}_i^T\mathbf{v}_j)}, \quad k = 1,\ldots,K-1, \tag{4}$$

where $\boldsymbol{v}_K = \mathbf{0}$ for identifiability and $\sum_{k=1}^{K}\pi_{ik} = 1$. Thus, the functional clustering model can be written as

$$\begin{aligned} &\mathbf{y}_i\,|_{z_{ik}=1} = \mathbf{S}_i(\boldsymbol{\mu}_k + \boldsymbol{\gamma}_i) + \boldsymbol{\varepsilon}_i,\ i = 1,\ldots,n, \\ &\boldsymbol{\gamma}_i \sim N(\mathbf{0},\boldsymbol{\Gamma}_k), \quad \boldsymbol{\varepsilon}_i \sim N(\mathbf{0},\sigma^2\mathbf{I}_{n_i}), \end{aligned} \tag{5}$$

where $\boldsymbol{S}_i = (\boldsymbol{s}(t_{i1}),\ldots,\boldsymbol{s}(t_{in_i}))^T$ is the spline basis matrix for the i-th curve.

2.2. Parameter estimation

We recommend using the EM-algorithm to obtain the MLE of all the parameters. Since the $\boldsymbol{z}_i$'s and $\boldsymbol{\gamma}_i$'s are assumed independent each other, the combined density distribution of the complete data $\{\boldsymbol{y},\boldsymbol{\gamma},\boldsymbol{z}\}$ can be expressed as

$$p(\mathbf{y},\boldsymbol{\gamma},\mathbf{z}) = p(\mathbf{y}\,|\,\boldsymbol{\gamma},\mathbf{z})\,p(\boldsymbol{\gamma}\,|\,\mathbf{z})\,p(\mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{y}_i\,|\,\boldsymbol{\gamma}_i,\mathbf{z}_i)\,p(\boldsymbol{\gamma}_i\,|\,\mathbf{z}_i)\,p(\mathbf{z}_i), \tag{6}$$

where

$$p(\mathbf{y}_i \mid \boldsymbol{\gamma}_i, \mathbf{z}_i) = \prod_{k=1}^{K} \{(2\pi\sigma^2)^{-\frac{n_i}{2}} \exp[-\frac{1}{2\sigma^2} \| \mathbf{y}_i - \mathbf{S}_i(\boldsymbol{\mu}_k + \boldsymbol{\gamma}_i) \|^2]\}^{z_{ik}},$$

$$p(\boldsymbol{\gamma}_i \mid \mathbf{z}_i) = \prod_{k=1}^{K} \{(2\pi)^{-\frac{q}{2}} \mid \boldsymbol{\Gamma}_k \mid^{-\frac{1}{2}} \exp(-\frac{1}{2} \boldsymbol{\gamma}_i^T \boldsymbol{\Gamma}_k^{-1} \boldsymbol{\gamma}_i)\}^{z_{ik}},$$

$$p(\mathbf{z}_i) = \prod_{k=1}^{K} \pi_{ik}^{z_{ik}}.$$

Let $p_k(\boldsymbol{y}|\boldsymbol{\theta}_k;\sigma^2)$ be the density of the k-th cluster where $\boldsymbol{\theta}_k = \{\boldsymbol{\Gamma}_k, \boldsymbol{\mu}_k, \boldsymbol{\nu}_k\}$, $\Theta = \{\Theta_k | k = 1,2,\dots,K\}$ where $\Theta_k = \{\boldsymbol{\theta}_k, \sigma^2\}$. The penalized log-likelihood of the complete data is given by

$$\begin{aligned} L(\theta) = & \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \log(\pi_{ik}) \\ & -\frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left(\log \mid \boldsymbol{\Gamma}_k \mid + \boldsymbol{\gamma}_i^T \boldsymbol{\Gamma}_k^{-1} \boldsymbol{\gamma}_i \right) \\ & -\frac{1}{2} \sum_{i=1}^{n} (n_i \log\sigma^2 + \sigma^{-2} \sum_{k=1}^{K} z_{ik} \| \mathbf{y}_i - \mathbf{S}_i(\boldsymbol{\mu}_k + \boldsymbol{\gamma}_i) \|^2) \\ & +\lambda \sum_{k=1}^{K} \| \mathbf{D}_d \boldsymbol{\mu}_k \|^2, \end{aligned} \tag{7}$$

where $\boldsymbol{D}_d$ is a d-th order difference penalty matrix such that $\boldsymbol{D}_d \boldsymbol{\mu}_k = \Delta^d \boldsymbol{\mu}_k$ constructs a vector of d-th differences of $\boldsymbol{\mu}_k$, and λ is a nonnegative tuning parameter to control the degree of smoothness of the fitting curve.

The EM algorithm consists of iteratively maximizing the expected value of the log-likelihood (7) by giving $\boldsymbol{y}_i$ with respect to the parameters. Using the conditional property of multi-normal distribution, we have

$$\boldsymbol{\gamma}_i \mid \mathbf{y}_i, z_{ik} = 1 \sim N((\sigma^2 \boldsymbol{\Gamma}_k^{-1} + \mathbf{S}_i^T \mathbf{S}_i)^{-1} \mathbf{S}_i^T (\mathbf{y}_i - \mathbf{S}_i \boldsymbol{\mu}_k), (\boldsymbol{\Gamma}_k^{-1} + \mathbf{S}_i^T \mathbf{S}_i / \sigma^2)^{-1}). \tag{8}$$

The expected value of z_{ik} given $\boldsymbol{y}_i$ is

$$\pi_{k|i} = P(z_{ik} = 1 \mid \mathbf{y}_i) = \frac{\pi_{ik} p(\mathbf{y}_i \mid z_{ik} = 1)}{\sum_{j=1}^{K} \pi_{ij} p(\mathbf{y}_i \mid z_{ij} = 1)}. \tag{9}$$

The covariance matrix $\boldsymbol{\Gamma}_k$ can be calculated using that

$$\boldsymbol{\Gamma}_k = \frac{1}{\sum_{i=1}^{n} \pi_{k|i}} \sum_{i=1}^{n} \pi_{k|i} E[\boldsymbol{\gamma}_i \boldsymbol{\gamma}_i^T \mid \mathbf{y}_i, z_{ik} = 1]. \tag{10}$$

Then we maximize the expected value of (7) and obtain that

$$\boldsymbol{\mu}_k = (\sum_{i=1}^{n} \pi_{k|i} \mathbf{S}_i^T \mathbf{S}_i + \lambda \mathbf{D}_d^T \mathbf{D}_d)^{-1} \sum_{i=1}^{n} \pi_{k|i} \mathbf{S}_i^T (\mathbf{y}_i - \mathbf{S}_i \hat{\boldsymbol{\gamma}}_{ik}), \tag{11}$$

where $\hat{\gamma}_{ik} = E[\gamma_i | y_i, z_{ik} = 1]$ can be calculated by (8).

The final step is to set

$$\sigma^2 = \frac{1}{\sum_{i=1}^{n} n_i} [\lambda \sum_{k=1}^{K} \| \mathbf{D}_d \boldsymbol{\mu}_k \|^2 + \sum_{i=1}^{n} \sum_{k=1}^{K} \pi_{k|i} (\| \mathbf{y}_i - \mathbf{S}_i \boldsymbol{\mu}_k - \mathbf{S}_i \hat{\boldsymbol{\gamma}}_{ik} \|^2 + \mathbf{S}_i^T Cov[\boldsymbol{\gamma}_i \mid \mathbf{y}_i, z_{ik} = 1] \mathbf{S}_i)]. \tag{12}$$

2.3. Curve clustering and model selection

Most of the existing curve clustering methods are based on curve shape clustering. The functional data clustering method in this chapter is not only based on curve shape, but also considers the influence of covariates on responses. Assuming that the data has been generated by (5) with K clusters, the model can be fitted by the method described in the previous section. Denote the parameter set $\hat{\theta} = \{\hat{\Gamma}_k, \hat{\mu}_k, \hat{\nu}_k\}$ and $\boldsymbol{\nu}_K = \mathbf{0}$. Therefore, given the observation curve $\boldsymbol{y}^*$ and the covariate $\boldsymbol{w}^*$ corresponding to the curve, the posterior distribution of the latent variable $z^* = (z_1^*, \dots, z_K^*)^T$ is

$$P(z_k^* = 1 \mid \mathbf{y}^*) = \frac{\pi_{ik}^* p(\mathbf{y}^* \mid z_{ik} = 1)}{\sum_{j=1}^{K} \pi_{ij}^* p(\mathbf{y}^* \mid z_{ij} = 1)},$$

where

$$\pi_{ik}^* = \frac{\exp(\mathbf{w}_i^{*T} \mathbf{v}_k)}{1 + \sum_{j=1}^{K-1} \exp(\mathbf{w}_i^{*T} \mathbf{v}_j)},$$

when $k = k^*, k = 1,2,\ldots,K$, $P(z_k^* = 1|\boldsymbol{y}^*)$ takes the maximum value, then the curve is divided into the k^*-th cluster.

In this chapter, the parameter fitting is carried out by means of the basis function in the nonparametric method. The commonly used basis functions include the Fourier basis function and the B-spline basis function. Due to the good properties of the derivatives of the B-spline basis function, it is the first choice in this paper. The determination of the position of the node and the selection of the order of the basis function are two major problems faced when using the B-spline for curve fitting. This chapter selects the cubic spline basis function and fits the curve better by penalizing the base coefficient.

For the determination of the penalty coefficient, Chen et al. takes the penalty coefficient as a parameter and uses the likelihood function to estimate it [14]. This chapter uses traditional generalized cross-validation methods to determine the penalty coefficient.

Choosing the number of clusters is an important but difficult problem in cluster analysis, which is to determine the value of K. This chapter focuses on the study of functional clustering methods. Therefore, for the determination of the number of clusters, the traditional computationally simple BIC criterion is used for discriminating [15].

$$BIC = -2L(\hat{\boldsymbol{\theta}}) + Glog(N),$$

where $L(\widehat{\boldsymbol{\theta}})$ is the value of the log-likelihood function, G is the total number of all unknown parameters, and N is the sample size.

3. Simulation study

This section illustrates the practicality of functional clustering methods through simulation studies. The simulation process gives the correct category information in advance, and measures the clustering quality by comparing the clustering results with known criteria. This section uses two indicators to measure cluster quality. The first is the correct classification rate (cRate). cRate is defined as the maximum possible ratio of the correct classification object to the total number of objects to be aggregated. The correct clustering refers to the known category information. The second discriminant indicator is the adjusted Rand (aRand) indicator proposed by Hubert and Arabie [16], this indicator is a revised version of Rand index [17]. The value of Rand index is $[0,1]$. When the value is larger, the clustering result is more consistent with the real situation, indicating that the clustering result is more accurate and the purity within each class is higher. The Rand index is further adjusted by aRand, so that its expected value is 0 and the range is $[-1,1]$. A larger value means that the clustering result is more consistent with the real situation. The simulation generates 60 curves, each of which contains 30 nodes. t is equally distributed in $[0,1]$, and the observation value y_i is generated by

$$y_i(t) = \mathbf{s}_k^T(t)(\boldsymbol{\mu}_k + \boldsymbol{\gamma}_i) + \varepsilon_i(t), \tag{13}$$

where $k = 1,2,3$, ie category $K = 3$. In different classes, $s(t)$ select 4 order B-spline basis function, the number of nodes is set to 10. Let $\boldsymbol{\mu}_k \sim N(\mathbf{0}, \boldsymbol{I}_q)$, $\boldsymbol{\gamma}_i \sim N(\mathbf{0}, \boldsymbol{\Gamma}_k)$, where $\boldsymbol{\Gamma}_k = \sigma_k^2 \boldsymbol{I}_q$, $\boldsymbol{I}_q$ is the q dimension unit matrix, $\sigma_1^2 = 0.1, \sigma_2^2 = 0.2, \sigma_3^2 = 0.5$. The random error $\varepsilon_i(t) \sim N(0, \sigma_\varepsilon^2)$, $\sigma_\varepsilon^2 \sim U(0,0.8)$. In addition, the number of curves in different classes is determined by the logistic structure of the form (4). The priori information $\boldsymbol{w}$consists of the $\mathbf{1}$ vector and the one-dimensional column vector, which follows the uniform distribution on $[-1,1]$. The coefficient $v = ((0,2)^T, (0,1)^T, (0,0)^T)$. So, the three types of curves generated are shown in Fig. 1. Fig. 2 shows the clustering results using the functional data clustering model (newCFDA) described in this paper.

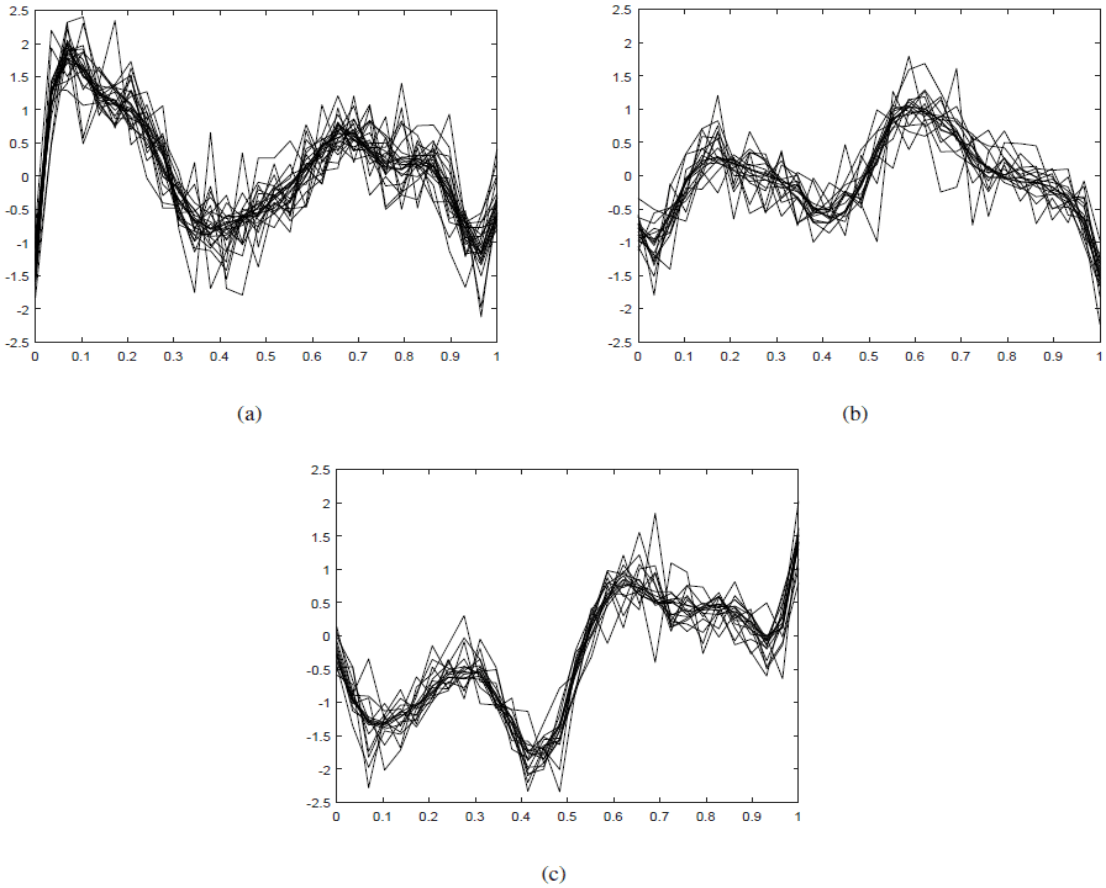

Fig. 1: Simulate the generation of 3 types of curve shapes.

Like James and Sugar [11], the low-rank expression shown in Fig. 2(b) uses the following formula (14) to convert $\boldsymbol{\mu}_k$ into the lower-dimensional α_k for drawing.

$$\boldsymbol{\mu}_k = \boldsymbol{\lambda}_0 + \boldsymbol{\Lambda}\boldsymbol{\alpha}_k, \tag{14}$$

where $\boldsymbol{\lambda}_0$ and $\boldsymbol{\alpha}_k$ are the q dimension vector and the h dimension vector respectively, $\boldsymbol{\Lambda}$ is a $q \times h$ matrix, $h \leq min(q, K-1)$, in this article, we take $h = 2$. This low rank expression makes it very straightforward to see that 60 curves are divided into 3 categories. The points of different shapes in the figure are the centralized projection of $\boldsymbol{y}_i$ on the h dimension space, and the solid origin is the estimated value of $\hat{\boldsymbol{\alpha}}_k$ for different categories.

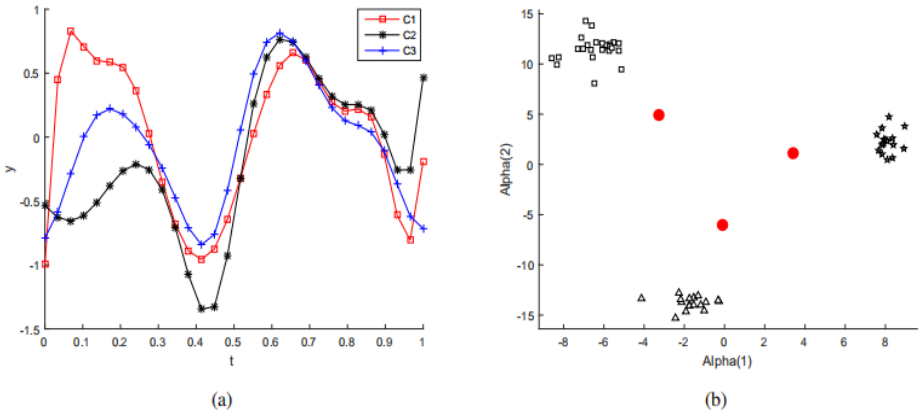

Fig. 2: Clustering result. (a) Various cluster center curves (b) Low rank expression

The generated simulation data is used for the functional clustering model (CFDA) of the sparse data [11], the new CFDA model described in this article, k-centres functional clustering model (KCFC) of Chiou and Li [12], functional principal component analysis (FPCA) [18] and K-means clustering method. Repeat 500 experiments and record the average cRate and aRand indicators of several methods. The results are shown in Table 1.

Table 1: Comparison of the accuracy of several clustering methods.

Model	aRand		cRate	
	mean	Standard deviation	mean	Standard deviation
New CFDA	0.90	0.08	0.94	0.06
CFDA	0.82	0.12	0.90	0.08
KCFC	0.47	0.23	0.41	0.38
FPCA	0.81	0.06	0.87	0.05
K-means	0.71	0.12	0.83	0.08

It is not difficult to see from Table 1. For comparison, for the three model based functional clustering methods, the improved functional clustering method (newCFDA) has the best clustering result, and the average accuracy rate is reached 94%, and the fluctuation is relatively stable, the functional clustering method (CFDA) of sparse data is second, the accuracy rate is 90%, and the k-centres functional clustering method has a poor clustering effect on the simulated data, and the error rate is higher, the fluctuation range is relatively large, relatively unstable, and the other two are not based on the model, the functional principal component analysis method that only clusters from the data has an average accuracy of 87%, which is relatively stable, and based on the K-means method of clustering discrete data is relatively inaccurate because it simply clusters the distance between vectors.

Through the above simulation analysis, it is not difficult to find that for the clustering study of functional data, adding valid prior information and considering the variance of the different types of data can effectively improve the clustering accuracy. The next section will examine the actual data using the new CFDA method improved in this paper.

4. Analysis of real data

The data used in this section comes from the website http://beijingair.sinaapp.com/. The air quality index (AQI) is calculated from the concentration values of the six pollutants PM2.5, PM10, SO2, O3, CO, NO2. This section selects the concentration of pollutants in 161 cities from March 1, 2015 to February 29, 2016. The research period is the four seasons of spring, summer, autumn and winter. Studies have shown that human factors have an inseparable impact on urban air quality. Urban green space coverage, industrial sulfur dioxide emissions and industrial dust emissions can reflect to some extent the degree of human impact on urban air quality. The relevant variables are selected from 2016 China Urban Statistical Yearbook.

According to the Air Quality Index, the Ministry of Environmental Protection divides China's air quality into six grades, as shown in Table 2. Concerned about the air pollution situation in 161 cities, according to the air quality index of each city, the number of days of mild pollution, moderate pollution, severe pollution and particularly severe pollution appeared in the selected time interval. Draw as shown in Fig. 3.

Table 2: Air quality index levels and division criteria

Air quality index	Level	Color	Health impact
0˜50	Excellent	Green	No effect basically
51˜100	Well	Yellow	May have a weaker impact on the health of very few people
101˜150	Mild pollution	Orange	Symptoms in susceptible populations increase and irritations in healthy people
151˜200	Moderate pollution	Red	Further aggravating the symptoms of susceptible people, may affect the heart, respiratory system of healthy people
201˜300	Severe pollution	Purple	Heart disease, lung disease patients with symptoms, and symptoms in healthy people
> 300	Particularly severe pollution	Maroon	Exercise tolerance in healthy people is reduced and certain diseases appear in advance

It is not difficult to see from the Fig. 3, in Northeast China and North China, such as Shanxi and Hebei, there are more days of pollution in the year, and there are fewer times of pollution in southern China and Southwest China. Overall, the air quality in southern China is better than that in the north and the West is better than the East. This also shows that the development of a city's economy and the type of pillar industry have a certain degree of impact on its air quality. The concentration of PM2.5 in the dusty air emitted by industrial production has become one of the main pollutants affecting China's air quality. Table 3 is the air quality index level published by the environmental protection department according to the daily average PM2.5 concentration range.

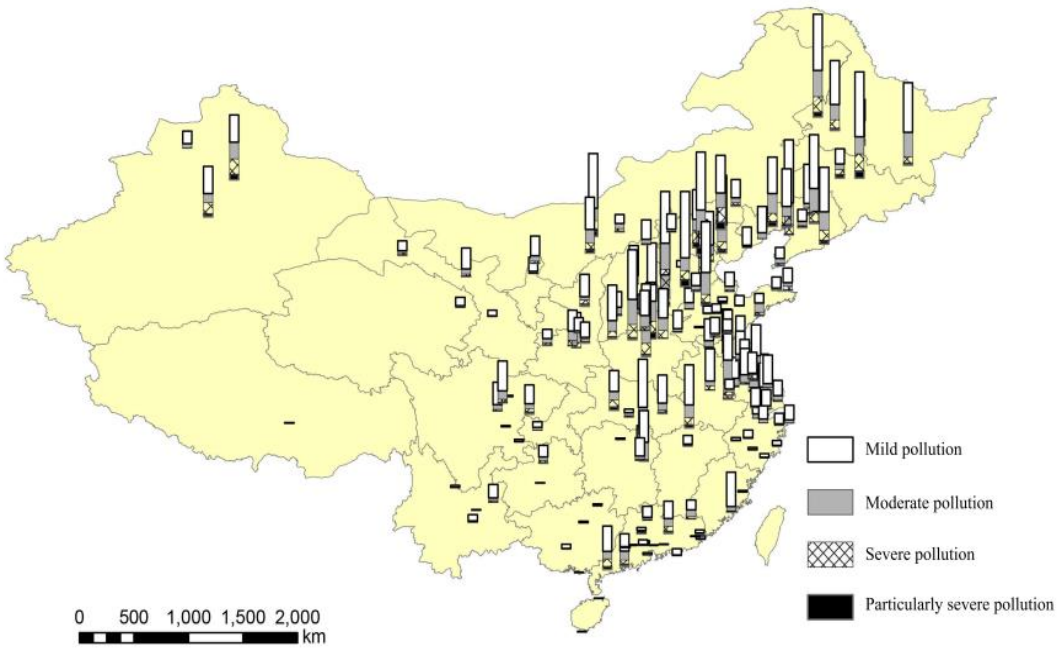

Fig. 3. Air pollution in 161 cities in 2015

Table 3: $PM_{2.5}$ concentration and air quality index level division criteria

concentration ($\mu g/m^3$)	level
0~35	Excellent
35~75	Well
75~115	Mild pollution
115~150	Moderate pollution
150~250	Severe pollution
250~500	Particularly severe pollution

In this section, we use the improved functional data clustering method to cluster the daily average PM2.5 concentration values of 161 cities in China in the spring, summer, autumn and winter seasons of 2015-2016. Based on the above clustering method, considering that different categories should have large differences, this section is based on the functional data clustering method of sparse data proposed by James and Sugar (2003), let different categories have different variances. In addition, the urban air quality status is affected to some extent by human factors. Therefore, we use urban latitude and longitude, urban green area coverage, industrial sulfur dioxide emissions and industrial smoke dust emissions as priori information of clustering. When we fit the curve with the B-spline, we add a partial penalty factor to the base coefficient, and adjust the use of the basis function according to the penalty parameter, which further optimizes the fitting result of the B-spline.

Fig. 4 shows the daily average PM2.5 concentration of 161 cities in China in the selected time range. From the figure, we can only roughly see that the concentration of PM2.5 of each city have peaks in winter, but it is difficult to understand the situation of each city. We use the functional clustering model presented in this paper to cluster the PM2.5 concentration in the selected city, and obtain the clustering results as shown in Fig.5.

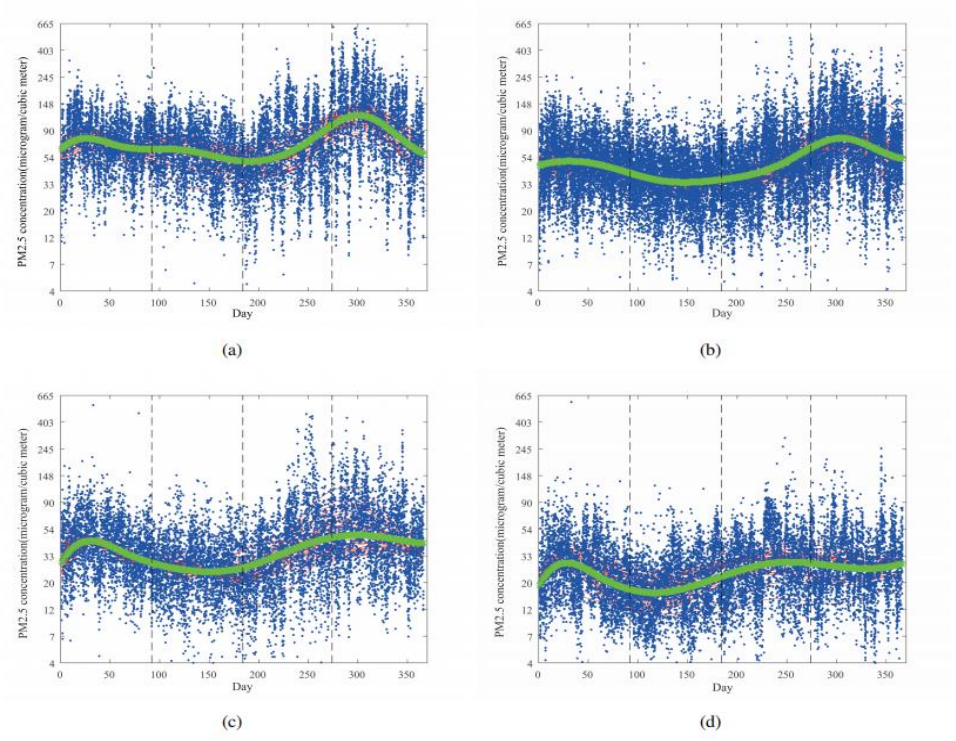

Fig. 4: Daily observation $PM_{2.5}$ concentration of 161 cities in 2015

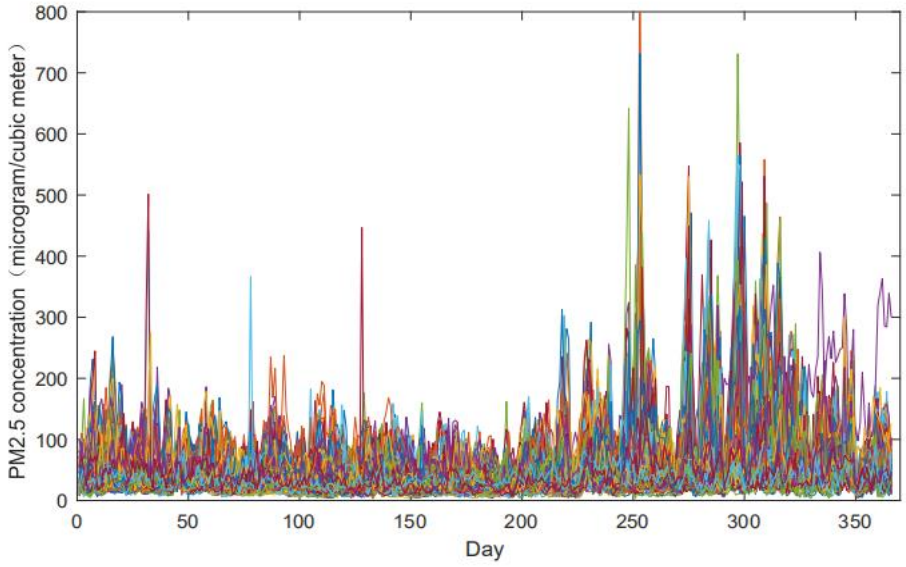

Fig. 5: Clustering results of $PM_{2.5}$ concentration in 161 cities. Discrete points represent $PM_{2.5}$ concentration values, the thin line is the fitted curve for the $PM_{2.5}$ concentration in each city, the thick line is the cluster center curve for each type of city.

The clustering method clusters 161 cities into four categories, among which 34 cities such as Shijiazhuang, Tangshan and Zhengzhou are grouped into one category (Fig. 5(a)), the average PM2.5 concentration is 75 micrograms per cubic meter, and the air quality has reached the pollution level. In many cities, the concentration of PM2.5 exceeds 500 micrograms per cubic meter, and even reaches 750 micrograms per cubic meter in winter. Then, 62 cities such as Beijing, Taiyuan, Shanghai and Nanjing are clustered into one category, as shown in the Fig. 5(b), with an average PM2.5 concentration of 51 micrograms per cubic meter, the PM2.5 concentration values of such cities are basically below 500 micrograms per cubic meter. Fig. 5(c) indicates that the 30 cities of Zhangjiakou, Guiyang and Kunming are grouped into one category, and the average PM2.5 concentration is 36 micrograms per cubic meter. Finally, in Fig. 5(d), 35 cities such as Lhasa, Zhangjiajie and Sanya are clustered into one class, and the average PM2.5 concentration is 25 micrograms per cubic meter, the PM2.5 concentration values of these cities are basically no more than 200 micrograms per cubic meter, which is a class of cities with better air quality. Overall, the proportion of cities with better air quality in the selected 161 cities is about 22%. Most of the cities have poor air quality and it is urgent to improve the environmental quality.

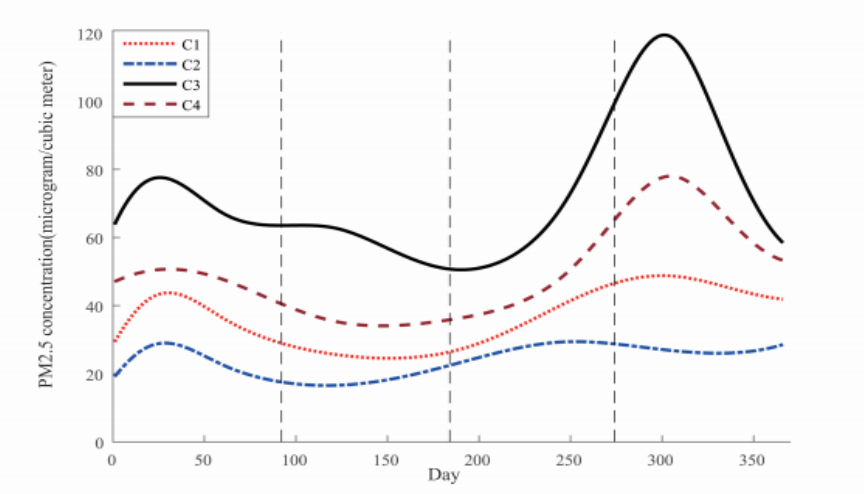

Fig. 6: PM2.5 concentration of urban cluster centre curve

In order to further analyze the time distribution of PM2.5 concentration in various cities, Fig. 6 shows the cluster center curve of various cities, the selected time interval is divided into spring, summer, autumn and winter seasons by vertical dashed lines.

It can be seen from the Fig. 6 that the concentration fluctuations of different categories of PM2.5 are different. From the top to the bottom, in the first class of cities, the PM2.5 concentration peaked in the spring, and then showed a slow decline trend, which began to decrease significantly in the summer, and the valleys appeared in the early autumn. Finally, an apparent peak appears in winter and then drop sharply. The concentration of PM2.5 in the second category of cities showed peaks in spring and winter, and the peak was significantly higher in winter than in spring. For the third type of city, there are also two peaks, but the second peak appears earlier than the second category of cities, almost at the time of the autumn and winter handover, and the trough appears in the late summer. Different from the second and third categories of cities, the second peak of the PM2.5 concentration in the fourth category of cities appears in the autumn, the winter begins to decline slowly, and begins to rise again at the end of winter, and its trough period appears in the early summer.

For the trend of PM2.5 concentration change with seasons in different types of cities, relevant departments can strengthen prevention before the concentration rises according to their own situation, and do a good job of precaution to delay the occurrence of the peak or control the peak value.

5. Conclusion

Clustering analysis is an effective method to simplify the data structure. Clustering analysis of functional data is more difficult than ordinary discrete data. Model-based clustering analysis is a widely used method. The functional data clustering method used in this paper is applicable to various functional data, even functional data with non-homogeneous or sparse time observations.

Based on the functional clustering method of this paper, the simulation research is carried out. Both graph and numerical results show the practicability of the method, and it has unique advantages compared with other model-based clustering methods and data-based clustering methods. We use this clustering method in the cluster analysis of PM2.5 concentration in 161 cities in China, and divide the selected cities into four different categories. The analysis found that the concentration of PM2.5 in various cities showed different evolution trends with seasonal changes, which were related to various factors such as climate factors, economic situation, pillar industries and government policies. Finally, through cluster analysis to study the spatial and temporal trends of various cities PM2.5 concentration can provide some effective suggestions for the government governance environment to some extent.

References

[1] J. Bao, X. Yang, Z. Zhao, et al. The spatial-temporal characteristics of air pollution in China from 2001-2014. International Journal of Environmental Research and Public Health, 2015, 12(12): 15875-15887.

[2] H. Pu, K. Luo, P. Wang, et al. Spatial variation of air quality index and urban driving factors linkages: evidence from Chinese cities. Environmental Science and Pollution Research, 2017, 24(5): 4457-4468.

[3] Y. Chen, L. Wang, F. Li, B. Du, K. K. R. Choo, H. Hassan, and W. Qin. Air quality data clustering using EPLS method. Information Fusion, 2017, 36: 225-232.

[4] A. A. Hamedian, A. Javid, S. M. Zarandi, Y. Rashidi, and M. Majlesi. Air quality analysis by using fuzzy inference system and fuzzy c-mean clustering in Tehran, Iran from 2009-2013. Iranian Journal of Public Health, 2016, 45(7): 917.

[5] J. O. Ramsay, B. W. Silverman. Functional data analysis, 2nd ed. New York: Springer, 2005.

[6] E. A. Maharaj. Cluster of time series. Journal of Classification, 2000, 17(2): 297-314.

[7] C. Iorio, G. Frasso, A. DAmbrosio, et al. Parsimonious time series clustering using p-splines. Expert Systems with Applications, 2016, 52: 26-38.

[8] A. Sam, F. Chamroukhi, G. Govaert, et al. Model-based clustering and segmentation of time series with changes in regime. Advances in Data Analysis and Classification, 2011, 5(4): 301-321.

[9] C. Bouveyron, J. Jacques. Model-based clustering of time series in group-specific functional subspaces. Advances

in Data Analysis and Classification, 2011, 5(4): 281-300.

[10] S. Shoham. Robust clustering by deterministic agglomeration EM of mixtures of multivariate t-distributions. Pattern Recognition, 2002, 35(5): 1127-1142.

[11] G. M. James, C. A. Sugar. Clustering for sparsely sampled functional data. Journal of the American Statistical Association, 2003, 98(462): 397-408.

[12] J. M. Chiou, P. L. Li. Functional clustering and identifying substructures of longitudinal data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2007, 69(4): 679-699.

[13] J. Q. Shi, B. Wang. Curve prediction and clustering with mixtures of Gaussian process functional regression models. Statistics and Computing, 2008, 18(3): 267-283.

[14] H. Chen, Y. Wang. A penalized spline approach to functional mixed effects model analysis. Biometrics, 2011, 67(3): 861-870.

[15] G. Schwarz. Estimating the dimension of a model. The Annals of Statistics, 1978, 6(2): 461-464.

[16] L. Hubert, P. Arabie. Comparing partitions. Journal of Classification, 1985, 2(1): 193-218.

[17] W. M. Rand. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association, 1971, 66(336): 846-850.

[18] V. Zipunnikov, B. Caffo, D. M. Yousem, et al. Multilevel functional principal component analysis for high-dimensional data. Journal of Computational and Graphical Statistics, 2011, 20(4): 852-873.