Functional Data Clustering Method Based on Shape Information and Functional Mahalanobis Distance

Zibing Wang*

School of economics, Henan University, Kaifeng, 475001, China

*Corresponding author: neve_w@163.com

Abstract. Function-type data can provide insights into the internal structure of the data and facilitate the extraction of data features from the perspective of interactive derivative functions. The paper proposes a nonparametric clustering method for function-based data that incorporates first-order and second-order derivative function information into the Marginal distance for clustering function-based data. The method is based on the traditional K-means algorithm and is designed to cluster function-based data that contains rich shape information. The paper evaluated the performance of the algorithm of the text by comparing its purity and adjusted Rand coefficients against six other clustering algorithms on three different datasets. The results show that the algorithm of the text outperforms the other algorithms, demonstrating its outstanding performance, wide applicability, and practical significance in solving real-world problems.

Keywords: Function-based data, Clustering algorithm, Nonparametric clustering, Derivative function information

1. Introduction

In this age of information abundance, data no longer exist as fragmented and sparse pieces, but rather as continuous dynamic functions. For instance, in financial markets, stock transaction prices, electrocardiogram data, the growth curve of the human body, and real-time traffic flow in traffic control, all take values continuously within a certain space-time interval and exhibit a significant functional trend. This type of data is called functional data [1]. Along with functional data comes the emergence of functional data analysis methods. Functional data clustering analysis is an important research direction in the field of data mining and artificial intelligence. For instance, clustering stock trading prices to identify the trend characteristics of stocks, clustering weather data to identify the spatial and temporal characteristics of climate, and clustering macroeconomic indicator data to identify the economic characteristics among different countries. Especially in the field of biology, functional data clustering has wide and deep applications, which are significant for gene expression and function studies.[2]

There are three main function-based clustering methods: two-stage clustering, model-based clustering, and nonparametric clustering. Two-stage clustering involves smoothing and reducing infinite-dimensional functional data to a finite coefficient space and then applying traditional clustering methods to the finite-dimensional coefficients. For instance, Giaocofcien[3] and other researchers used clustering based on wavelet basis function. Wu Qiran et al. applies K-Means clustering of principal component basis coefficients of air quality data to explore the characteristics of air quality changes in Zhejiang Province[4]. However, two-stage clustering depends heavily on the number and type of basis functions, and different basis functions can lead to contradictory results. Model-based clustering methods involve clustering based on probabilistic methods, and designing clustering methods based on the probability distribution of the base expansion coefficient vector. For instance, Gao minghui and other researchers concluded that the adaptive clustering method is suitable for data with missing problem and clustered the time-observed data of traditional Chinese medicine with the feature to mine population characteristics[5]. Faicel and other reaserchers present model-based clustering approaches along with efficient algorithmic tools for clustering and classifying functional data[6]. However, the accuracy of probability density estimation may be affected due to insufficient samples in the group, reducing the efficiency of clustering. Non-parametric clustering

methods are clustering methods based on similarity measures between random functions. For example, Chiou [7] et al. proposed a clustering method for estimating class means and covariances by nonparametric methods. For more researches about non-parametric clustering methods, please refer to: [9][8].

Functional data allows the intrinsic features of the data to be mined from the perspective of the interacting derivative functions. However, most functional clustering methods currently in use overlook the derivative information in data. Junpeng Guo[10] and other researchers have developed hierarchical clustering algorithms that incorporate both the actual function distance and the first-order derivative function distance to ensure that clustering results are not only close in terms of distance but also have similar morphological features. However, for data samples with similar function trends and consistent fluctuation ranges, the first-order derivative function is insufficient to recognize key features.

This paper proposes using both first-order and second-order derivative function information in the clustering algorithm for function-based data. For data smoothing, functional principal component analysis is used to approximate discrete samples and reduce data noise. Specifically, the approach involves using functional principal component analysis to extract first-order and second-order derivative function information and form the derivative eigenvector. The Mahalanobis distance is then used to calculate similarity between different samples, and K-mean clustering algorithm is used to complete functional clustering. Actual data analysis shows that the proposed method outperforms most comparative methods.

2. Theory and methodology

2.1. Functional Principal Component Analysis

Functional data refers to data that is represented in the form of a function, where the entire function is considered as a singular unit of data. However, in reality, the values obtained from observations are often discrete points, and there may be errors in the observations. Therefore, it is necessary to reconstruct the actual functions that are hidden behind the observation data.[11]

Suppose that $X_i(t)$ (i=1,2,…n) is the object of function under study, $y_{i1}, y_{i2}, \ldots, y_{iT_i}$ are the T_i observations of function $X_i(t)$, $\varepsilon_{ij} (j = 1,2, \ldots T_i)$ is the observation error of the $j-th$ observation value of object ,which leads to quation(1):

$$y_{ij} = x_i(t_j) + \varepsilon_{ij} \tag{1}$$

For the estimation of $X_i(t)$, expand it over a set of basis functions $\Phi(t) = \{\phi_1(t), \phi_2(t), \ldots, \phi_k(t)\}'$:

$$X_i(t) = \sum_{k=1}^{K} \phi_k(t) c_k^i \tag{2}$$

Let $\boldsymbol{c}_i = (c_{i1}, c_{i2}, \ldots, c_{ik})'$, represent Eq.(2) in the matrix form:

$$x_i(t) = \Phi(t)' \boldsymbol{c}_i \tag{3}$$

Substituting Eq.(2) into Eq. (1), the base expansion coefficients are estimated according to the least square rule:

$$\boldsymbol{c}_i = argmin \sum_{j=1}^{T_i} \left(y_{ij} - \sum_{k=1}^{K} \phi_k(t) c_k^i\right) \tag{4}$$

Similarly, the first-order and second-order derivative of $x_i(t)$ are estimated:

$$x_i'(t) = \Phi(t)' \boldsymbol{a}_i \tag{5}$$

$$x_i''(t) = \Phi(t)' \boldsymbol{b}_i \tag{6}$$

Where the $\boldsymbol{a}_i$ and $\boldsymbol{b}_i$ are base-expanded coefficient vectors of $x_i'(t)$ and $x_i''(t)$ respectively. After smoothing the functions by basis expansion, the paper extracts K-dimensional vectors to

represent the original samples. However, considering the text utilizing both first-order and second-order derivative information in the clustering distance, 2k-dimension features are extracted for each sample[4] [11] [12][12] [13]. In order to improve computing efficiency and reduce data noise, the text applies functional principal component analysis to reconstruct the samples. To start with, let $N = n - 1$, the sample mean function and sample covariance function are used to estimate the mean function $\mu(t)$ and covariance function $V(s,t)$:

$$\hat{\mu}(t) = \frac{1}{N}\sum_{i=1}^{n} x_i(t) \tag{7}$$

$$\hat{V}(s,t) = \frac{1}{N}\sum_{i=1}^{n}(x_i(s) - \hat{\mu}(s))(x_i(t) - \hat{\mu}(t)) \tag{8}$$

The covariance function is then spectrally analyzed to obtain the characteristic equation:

$$\int_0^T f_j(t)\, V(s,t)dt = \sum_{j\geq 1} \lambda_j f_j(s) \tag{9}$$

Where $\lambda_1 \geq \lambda_2 \geq \ldots$,and when j=j' $\int_0^T f_j(t)\, f_{j'}(t)dt=1$;when j≠j' , $\int_0^T f_j(t)\, f_{j'}(t)dt=1$. Besides, $f_j(t)$ denotes eigenfunction and λ_j denotes eigenvalue. The basis function expansion of the characteristic function is given as $f(t)=\sum_{k=1}^{K} \phi_k(t)b_k$ i.e. $f(t)=\mathbf{b}\Phi(t)'$,where $\mathbf{b}=(b_1,b_2,\ldots,b_k)$is the parameter to be estimated. Meanwhile, substituting Eq.（3）into Eq. (7）yields:

$$\hat{V}(s,t) = N^{-1}X'X = N^{-1}\Phi(s)'\boldsymbol{c}'\boldsymbol{c}\Phi(t) \tag{10}$$

Replace V(s,t) by $\hat{V}$(s,t), and substitute the above two equations into the characteristic equation:

$$\int_0^T f_j(t)\, \hat{V}(s,t)dt=\int N^{-1}\Phi(s)'\mathbf{c}'\boldsymbol{c}\Phi(t)\Phi(t)'\boldsymbol{b}dt=N^{-1}\Phi(s)'\mathbf{c}'\boldsymbol{c}\mathrm{W}\boldsymbol{b} \tag{11}$$

$$\sum_{j\geq 1} \lambda_j f_j(s) = \boldsymbol{\rho}\boldsymbol{b}\Phi(s)' \tag{12}$$

Where $\mathrm{W} = \int \Phi(t)\Phi(t)'\, dt, \boldsymbol{\rho} = (\lambda_1, \lambda_2, \ldots)$. Organize the above equations and Eq.(11) can be simplified as:

$$N^{-1}\boldsymbol{c}'\boldsymbol{c}W\boldsymbol{b} = \boldsymbol{\rho}\boldsymbol{b} \tag{13}$$

which satisfy $\|f_j\| = 1$, i.e. $\mathbf{b}'\mathrm{W}\mathbf{b}=\mathbf{1}$. If $\{f_1,f_2,\ldots,f_L\}$is a set of orthogonal basis functions, then by the $K - L$ Expansion Theorem, $x_i(t)$ can be expanded as:

$$x_i(t) \approx \widehat{x_i}(t) = \mu(t) + \sum_{m=1}^{L} a_{im} f_m(t)\,, L \leq K \tag{14}$$

where a_{im} stands for the main component scores:

$$a_{im} = \int x_i(s) f_m(s) ds \tag{15}$$

The global approximation criterion is to minimize the following objective function:

$$PCASSE = \sum_{i=1}^{N} \|x_i - \hat{x}_i\|^2 = \int_0^T [x(t) - \hat{x}(t)]^2 dt \tag{16}$$

Similarly, the principal component expansion of $x_i'(t)$ and $x_i''(t)$ can be obtained as follows:

$$x_i'(t) \approx \widehat{x_i'}(t) = \mu'(t) + \sum_{m=1}^{L} a'_{im} f'_m(t)\,, L \leq K \tag{17}$$

$$x_i''(t) \approx \widehat{x_i''}(t) = \mu''(t) + \sum_{m=1}^{L} a''_{im} f''_m(t)\,, L \leq K \tag{18}$$

Their eigenvalues are respectively $\boldsymbol{\rho}'=(\lambda'_1,\lambda'_2,\ldots)$, $\boldsymbol{\rho}''=(\lambda''_1,\lambda''_2,\ldots)$ and their principal component score vectors are $\mathbf{a}'_{\mathrm{im}}=(a'_{i1},a'_{i2},\ldots,a'_{iL})'$; $\mathbf{a}''_{\mathrm{im}}=(a''_{i1},a''_{i2},\ldots,a''_{iL})'$. So far, each sample is downscaled from 2K features to 2L.

2.2. K-means algorithm based on derivative information and mahalanobis distance (DMk_means)

Based on the idea of simultaneously incorporating the first-order derivative function and second-order derivative function information of function-type data into the distance for clustering, this paper uses the first-order derivative function and second-order derivative function principal component scores of function-type data to be put into the vector $\boldsymbol{m_i}$. Meanwhile, the data is normalized by using the Mahalanobis distance. Represent the information of $X_i(t)$ by $\boldsymbol{m_i} = \left[\boldsymbol{a'}_{im}{}', \boldsymbol{a''}_{im}{}'\right]$. Denote that $\boldsymbol{\lambda} = [\ \boldsymbol{\rho'}, \boldsymbol{\rho''}]$. Define the distance as:

$$d\left(X_i(t), X_j(t)\right) = \left(\sum_{k=1}^{2L} \frac{(m_i^k - m_j^k)^2}{\lambda_k}\right)^{\frac{1}{2}} \tag{19}$$

The text adopts the K-Means algorithm to perform clustering. [14]The core idea of K-Means clustering is: first, set the target number of classes k for clustering and randomly classify the sample points into k classes; then, calculate the centroid of each class and the distance from each sample point to the k centroids, and reclassify the sample points into the classes closest to the class centroids. Finally, the previous step is repeated until the class centroids converge (i.e., the movement of class centroids tends to be stationary). The specific steps are as follows:

① Determine the number of clusters k, input $n$sample data $\mathbf{m_i}$, classify the samples into$k$ classes randomly, thus getting k set of classes K_h, and let the number of iterations be s (s=0,1,2,…)

② Compute class centroid $\bar{\mathbf{X}}_{\mathbf{K_h}}^{\mathbf{S}} = \frac{1}{K_h}\mathbf{m}_{\mathbf{K_h}}$,(h=1,2,…,k) . $\bar{\mathbf{X}}_{\mathbf{K_h}}^{\mathbf{S}}$ denotes the class centroid of class K_h after the $s-$th iteration.

③ Calculate the distance of each sample to the center point of the class according to equation (7). The samples are included in the set, the centroid of which they have the smallest distance value with.

④ Repeat step ② and ③ until the class centroid converges (increasing the number of iterations by one per repetition), which means: for $\varepsilon \in \boldsymbol{R}, \sum_{h=1}^{k}\left\|\bar{\boldsymbol{X}}_{\boldsymbol{K_h}}^{\boldsymbol{S}} - \bar{\boldsymbol{X}}_{\boldsymbol{K_h}}^{\boldsymbol{S-1}}\right\| \leq \varepsilon\ , (s \geq 1)$. At last, output the class center point $\bar{\boldsymbol{X}}_{\boldsymbol{K_h}}^{\boldsymbol{S}}$and the set of class K_h.

3. Data analysis

3.1. Experimental setup and data sources

This paper focuses on the experiments done on two datasets from the UCR dataset and one dataset from R, namely ECG200 and BeetleFly datasets (http://www.timeseriesclassification.com) and MedFly dataset (https://fda.readthedocs.io/en/latest/modules/datasets.html). The text used R for the experiments. The ECG200 dataset records ECG data from 200 patients. It contains 96 discrete observations of the ECG curve and is divided into two groups: normal heartbeat and myocardial infarction. ECG data is used in the medical field to determine the health of the heart. Patients with acute myocardial infarction often show pathologic Q-waves, ST-segment elevation, and T-wave inversion in their ECG. However, heart diseases are complex, and ECG data is variable, which requires experienced doctors to give accurate diagnosis. To assist doctors in the initial diagnosis and improve medical efficiency, clustering algorithms can identify the significant differences between various cardiac states. The BeetleFly dataset (later referred to as the Fly dataset) records centrifugal data of 20 sets of insect contours that are specifically categorized into two groups: bees and flies. Each data set contains 513 discrete observations. The text tests the validity of the proposed method by examining whether it can recognize significant differences between different graphical data. The MedFly dataset records 534 samples of egg-laying data of Mediterranean medflies surviving more than 34 days. The dataset labels medflies into two categories of long-lived and short-lived according to the time they survive. Specifically, it set the threshold value as 40 days, when the medfly survives

more than 40 days, the medfly is judged as a long-lived medfly, and vice versa as a short-lived medfly. The dataset includes 534 samples, and each sample takes the first 31 observation time points as predictor variables. By clustering the MedFly dataset, this paper can explore the relationship between Drosophila egg-laying patterns and longevity. The dataset information is shown in Table1 and the samples are shown in Figure 1.

Table 1. Dataset introduction

Data set	Information background	Sample	Dimension	Class
ECG200	Electrocardiographic data of normal and myocardial infraction	100	97	2
Fly	Centrifugal data of contour images of bees and flies	20	513	2
Medfly	medflies surviving more than 34 days	534	31	2

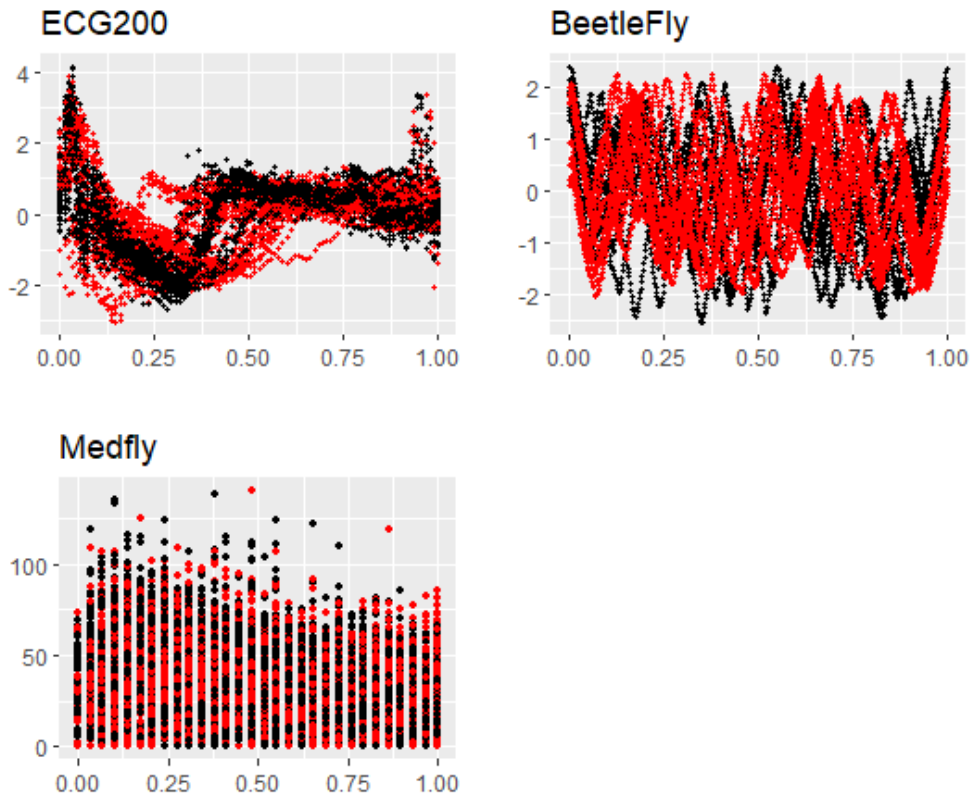

Figure 1. samples of the dataset

To detect the algorithm effect, this paper chose to use the Adjustment Rand Index (ARI) and Purity (PURTIY). ARI is utilized to measure the similarity between the clustering results and the real categories. It considers the effect of random assignment, and its value ranges from -1 to 1, where a larger value indicates better clustering results. The formula for ARI is as follows:

$$ARI = \frac{\Sigma_i\Sigma_j\binom{n_{ij}}{2} - \frac{\Sigma_i\binom{a_{ij}}{2}*\Sigma_j\binom{b_{ij}}{2}}{n/2}}{\frac{\Sigma_i\binom{a_{ij}}{2}+\Sigma_j\binom{b_{ij}}{2}}{2} - \frac{\Sigma_i\binom{a_{ij}}{2}*\Sigma_j\binom{b_{ij}}{2}}{n/2}} \tag{20}$$

Where, n_{ij} denotes the number of samples where the true label and the clustered label overlap (i denotes the true label and $j$denoteshe clustered label); a_{ij} denotes the total number of samples under the i-th class of true labels that overlap with the clustered labels, and , b_{ij} denotes the total number of samples under the j-th class of clustered labels that overlap with the true labels. The purity index is used to measure the accuracy of the clustered labels compared to the original labels.

3.2. Comparative results

In this paper, the ARI and PURITY indices obtained from this paper's algorithm are compared with their counterparts obtained from the waveclust[3], Itersubspace[7], distclust[15], funclust[16], funHDDC[17], and fscm[18] methods, respectively. The comparison results are shown in Table2:

Table 2. Algorithm test results

	ECG200		BeetleFly		MedFly	
	ARI	PURITY	ARI	PURITY	ARI	PURITY
DMk_means	0.2833	0.77	0.3244	0.8	0.0780	0.5674
distclust	0.0424	0.64	0.3304	0.8	0.0410	0.56
Itersubspace	0.2125	0.74	0.3304	0.8	0.0339	0.64
funclust	-0.0111	0.56	-0.0111	0.6	0.0231	0.55
funHDDC	0	0.64	-0.0398	0.55	0.0024	0.45
fscm	0.1031	0.67	0.4422	0.65	0.0346	0.49
waveclust	0	0.64	0.0526	0.65	0.1460	0.61

For the dataset ECG200, this paper's algorithm showed better performance in terms of ARI and purity index than the other six algorithms, indicating precise clustering performance. For the BeetleFly dataset, this paper's algorithm achieved a purity index of 0.8, which is the same as and better than the other four algorithms, while distclust, fscm, and Itersubspace algorithms showed better ARI index. In the case of the MedFly dataset, this paper's algorithm scored an ARI index of 0.07797999, which is lower than only waveclust, and a purity index of 0.5674157, which is lower than Itersubspace and waveclust methods. These results demonstrate that the algorithm presented in this paper is highly effective and has a wider range of applications.

4. Conclusion

This paper proposes a novel K-Means clustering method for function-type data that incorporates both the first-order and second-order derivative function information of the data, along with the function-type Mahalanobis distance. The aim is to cluster the data based on its intrinsic features. The effectiveness of this algorithm is demonstrated through experiments conducted on electrocardiogram data, insect contour centrifugal data and medflies data. The results indicate that the proposed method can effectively cluster the datasets and extract their intrinsic features. For the future, there is a lot of potential for development in the clustering method that uses information from the derivative function. First, the information of the derivative function can be directly used in two-stage clustering methods. The two-stage clustering method can maximize the retention of the derivative function information of the function-type data curves, and maximize the features of the sample based on the derivative function information.

References

[1] Wang J L, Chiou J M, Müller H G. Functional data analysis[J]. Annual Review of Statistics and its application, 2016, 3: 257-295.

[2] Zhang S, Li X, Lin J, et al. Review of single-cell RNA-seq data clustering for cell-type identification and characterization[J]. RNA, 2023, 29(5): 517-530.

[3] M., Giacofci, S., Lambert-Lacroix, G., & Marot, et al. (2013). Wavelet-based clustering for mixed-effects functional models in high dimension. Biometrics Journal of the Biometric Society An International Society Devoted to the Mathematical & Statistical Aspects of Biology.

[4] WU Qi-ran, Zhou Li-kai, SUN Jin-jin, et al. Researchon characeristics of air quality change in Zhejiang Province——based on functional data analysis[J]. Journal of Shandong University(Natural Science),2021,56(07):53-64.

[5] Gao Minghui, Yi Danhu, Peng jin, et al. Application of Functional Data Clustering Methods on Missing Data[J].Modernization of Traditonal Chinese Medicine and Materia Medica-World Science and Technology.2017,19(12):1966-1975.

[6] Chamroukhi F, Nguyen H D. Model-based clustering and classification of functional data[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2019, 9(4): e1298.

[7] CHIOU J M ,LI P L. Functional clustering and identifying substructures of longitudinal data[J]. Journal of the Royal Statistical Society Series B, 2007,69: 679-699.

[8] Zambom A Z, Collazos J A A, Dias R. Functional data clustering via hypothesis testing k-means[J]. Computational Statistics, 2019, 34: 527-549.

[9] Wu R, Wang B, Xu A. Functional data clustering using principal curve methods[J]. Communications in Statistics-Theory and Methods, 2022, 51(20): 7264-7283.

[10] GUO Jun-peng, WANG Mei-nan, GAO Cheng-ju, DAI Hui. Step-by-Step Hierarchical Algorithm for Functional Data [J].Journal of Systems & Management,2015,24(06):814-820.

[11] SUN Li-rong, ZHUO Wei-jie, WANG Kai-li, ma Jia-hui. Study on functional clustering analysis methods[J]. Applied Mathematics A Journal of Chinese Universities(Ser .A),2020,35(02):127-140.

[12] WANG DE-qing, ZHU Jian-ping, Liu Ciao-wei HE Ling-yu. Review and Prospect of Functional Data Clustering Analysis[J]. Journal of Applied Statistics and Management ,2018,37(01):51-63. 20170519-003.

[13] MENG Yinfeng, YANG Jiayu, CAO Fuyuan. Splist transfer hierarchical clustering algorithm for functional data[J]. Journal of Shandong University(Engineering Science),2022,52(01):19-27.

[14] LIANG B, LIANG J Y, CAO F Y. A multiple k-means clustering ensemble algorithm to find nonlinearly separable clusters[J]. Information Fusion,2020:61.

[15] Jie, Peng, Hans-Georg, & Müller. (2008). Distance-based clustering of sparsely observed stochastic processes, with applications to online auctions. The Annals of Applied Statistics.

[16] Jacques, J. , & Preda, C. . (2013). Funclust: a curves clustering method using functional random variables density approximation. Neurocomputing, 112(jul.18), 164-171.

[17] Charles Bouveyron and Julien and Jacques. Model-based clustering of time series in group-specific functional subspaces. Advances in Data Analysis and Classification.5(4).281-300.2011

[18] Nicoleta Serban and Huijing Jiang. Clustering Random Curves Under Spatial Interdependence With Application to Service Accessibility. Technometrics.54(2).108-119.2012