Noname manuscript No.
(will be inserted by the editor)

Unsupervised Learning on U.S. Weather Forecast Performance

Chuyuan Lin · Ying Yu · Lucas Y. Wu · Jiguo Cao

Received: date / Accepted: date

Abstract Nowadays, climate events and weather predictions have a huge impact on human activities. To understand the weather prediction accuracy, we applied the functional principal component analysis (FPCA) method to investigate the main pattern of variance within the U.S. weather prediction error over a period of 3 years. We further grouped the states in the U.S. based on their similarity in weather forecast performance using two types of functional clustering approaches: the filtering method and the model-based method. The strengths and shortages of each clustering method were detected through the simulation studies. Then, the clustering approaches were applied to U.S. weather data from 2014 to 2017. Through clustering, cluster-specific patterns were visually detected, and the cluster-to-cluster differences were quantified in order to identify the most and least predictable U.S. states.

Keywords Weather forecast · Functional principal component analysis · K-means clustering · discriminative functional mixture model · FunFEM model-based clustering

1 Introduction

Weather prediction affects human activities crucially, such as agricultural, fishery, industrial production and daily traveling [Adams et al. (1990)]. Accurate weather forecasts, especially short-term forecasts usually provide tremendous help and instruction to the preparations of the weather-sensitive industries and activities. Modern climatology and weather forecast techniques focus on predicting upcoming weather conditions, such as daily minimum temperature, based on current weather measurements. The weather forecast performance

Chuyuan Lin · Ying Yu · Lucas Y. Wu · Jiguo Cao (✉)
Department of Statistics and Actuarial Science, Simon Fraser University,
8888 University Drive, Burnaby, BC V5A 1S6, Canada
E-mail: jiguo_cao@sfu.ca

is usually evaluated by the prediction error: the difference between the actual and the forecast weather measurements [Bauer et al. (2015)]. Bauer and Thorpe suggested that the understanding of the climatic process and the input of statistical expertise are equally important to reduce the weather prediction error [Bauer et al. (2015)]. Then a big question arises to statisticians is how to improve the accuracy of weather forecasts based on statistical methods and modeling.

In the past decades, a number of different statistical methods were developed to forecast weather data. Two conventional approaches are parametric seasonal autoregressive integrated moving average (SARIMA) model [Box et al. (2015)] and non-parametric kernel predictor [Collomb (1983); Györfi et al. (1989); Bosq (1996)]. Due to the nature of weather data arriving in time, one can also apply functional data analysis, such as functional autoregressive (FAR) model [Besse et al. (2000)], which could produce an entire annual temperature trend one year ahead with a substantial reduction in mean square error compared to the traditional SARIMA model. With the increasing popularity of machine learning, support vector machine (SVM) and neural network were also applied to forecast weather data [Radhika and Shashi (2009)].

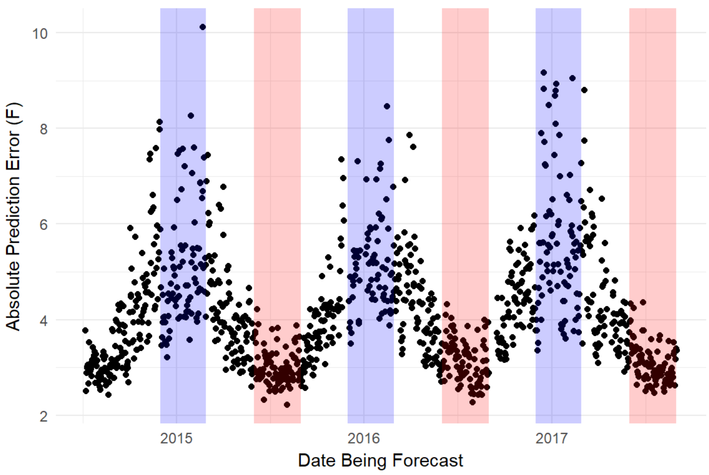


Fig. 1: Daily absolute prediction errors of minimum temperature change over time. The highlighted red regions represent summer periods and the blue highlighted regions represent winter periods.

Prediction errors in weather forecasting might be affected by the variation of weather conditions geographically or less accurate models used in weather forecasting [Orrell et al. (2001)]. Prediction errors may also vary depending on other variables, such as seasonality. Figure 1 shows the absolute daily prediction errors over a period of 3 years. The highlighted red and blue regions represent the summer and winter periods respectively. If the weather forecast is accurate, it is expected to see no clear patterns over time, where the errors look like a white noise. In Figure 1, it is clear to see the periodic pattern of prediction errors with a higher variance in the winter and lower variance in the summer time. Similarly, if the weather forecast is accurate, the errors should be similar across different geographical locations.

Figure 2 is a visualization mapping out the prediction accuracy of all 50 states in the U.S., where neighbouring states tend to have similar prediction accuracy (similar color in blue). Additionally, a similar climate condition may also contribute to the similarity of weather prediction accuracy, such as California and Hawaii. In this study, we referred this as the "spatio-climate effect", a term we created to address the joint effect of location and climate in weather forecasts.

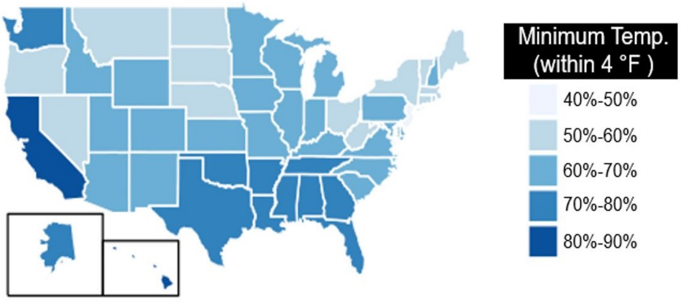


Fig. 2: Prediction accuracy of minimum temperature for each U.S. state. We consider the prediction is accurate if the prediction error is within 4 Fahrenheit, and the accuracy is evaluated as the percentage of days satisfying this condition within each state over a period of 3 years. A darker blue indicates a higher prediction accuracy.

Motivated by seasonal and spatio-climate patterns within weather forecast, the goals of our analysis are to focus on: 1). Explain how weather prediction performance changes over time; 2). Explore variations in weather forecast accuracy across different geographical locations in the U.S., and identify the most and least predictable regions.

Suggested by Ramsay and Silverman (2005), we reconstructed the daily prediction errors into functional from using smoothing B-spline. Then we applied functional principal component analysis (FPCA) on the smoothed data to obtain principal component functions, which can be used to study the overall patterns of variation within the prediction errors [Ramsay and Silverman

(2005), Luo et al. (2013), Lin et al. (2016), Sang et al. (2017)]. In literature, several clustering techniques were proposed for functional data, including filtering and model-based clustering methods [Jacques and Preda]. In the filtering approaches, instead of directly clustering on the fitted smoothed curves, a subset of feature information is extracted to represent the original data [James and Sugar (2003)]. For example, Abraham et al. (2003) and Adelfio et al. (2011) applied the K-means algorithm on the B-splines coefficients and functional principle component (FPC) scores. Ke et al. (2016) and Li et al. (2019) identified subgroups based on the homogeneity of regression coefficients. Model-based clustering methods approximate the cluster-specific probability distributions using the coefficients of the basis functions or FPC scores. One of the most recent model-based clustering algorithms is the FunFEM, which involves both the FPCA and discriminative functional subspace latent mixture models [Bouveyron et al. (2015)].

This paper has three major contributions. First, we investigate how the pattern of prediction errors of weather forecast changes over time for each state using the functional principal component analysis method. Second, we group similar states together via functional data clustering. Finally, we test three different clustering methods on simulated data to compare their performance under different simulation settings.

The paper is organized as follows. Methods are discussed in section 2. The application of the FPCA and functional clustering methods on real U.S. weather data is presented in section 3. The performances of the filtering and model-based clustering methods were compared through simulation studies in section 4. Section 5 concludes the paper.

2 Methods

2.1 Functional data reconstruction using B-spline

In practice, functional data is usually observed at discrete time points and the underlying functional expressions of the observed data are unknown. A common way to reconstruct data into the functional form is to express each observation in a finite dimensional space spanned by a set of basis functions. Consider a stochastic process $\{X(t) : t \in [a, b]\}$ at randomly located time points in $[a, b]$. The observations from n independent realizations of the process are

$$y_{ij} = X_i(t_{ij}) + \epsilon_{ij},$$

where y_{ij} is the j^{th} observation for i^{th} curve and $\{\epsilon_{ij}\}$ are i.i.d. random noise with mean 0 and variance σ^2. Then each $X_i(t)$ can be represented by a linear combination of Schoenbergs B-spline basis functions $\{\phi_p(t)\}_{p=1}^P$ as follows

$$X_i(t) = \sum_{p=1}^{P} \beta_{ip}\phi_p(t) = \boldsymbol{\beta}_i^T \Phi(t),$$

where $\boldsymbol{\beta}_i = (\beta_{i1}, \beta_{i2}, ...\beta_{iP})^T$ are the coefficients of the corresponding basis functions and can be estimated by least squares smoothing such that $\hat{\boldsymbol{\beta}}_i = \left[\Phi(t)^T\Phi(t)\right]^{-1}\Phi(t)^T y$ [Curry and Schoenberg (1966)]. Generalized cross validation (GCV) is used to choose the appropriate number of basis functions, P [Ramsay and Silverman (2005), Cao and Ramsay (2009)]. The optimal number is selected if it leads to the minimum of the GCV criterion

$$\text{GCV} = \frac{n^{-1}\text{SSE}}{[n^{-1}\text{trace}(I - S)]^2},$$

where SSE is the sum of squared errors and $S = \Phi(t)\left[\Phi(t)^T\Phi(t)\right]^{-1}\Phi(t)^T$ is the smoothing matrix [Ramsay and Silverman (2005)].

2.2 Smoothed functional principal component analysis

Smoothed functional principal component analysis (FPCA) is an extension of PCA with smoothness for investigating the dominate modes of variation of functional data. Given n functional curves $\{X_i(t) : t \in [a, b]\}_{i=1}^n$, we assume that they are i.i.d samples drawn from the distribution of a stochastic process $X(t)$ with covariance function $V(s, t) = E[(X(s)-E[X(s)])(X(t)-E[X(t)])]$, where s and t share the same domain $[a, b]$ [Ramsay et al. (2009)]. Using the Karhunen-Loeve decomposition [Ramsay et al. (2009)], $V(s, t)$ can be decomposed as

$$V(s, t) = \sum_{j=1}^{\infty} d_j\xi_j(s)\xi_j(t),$$

where d_j are the sorted eigenvalues such that $d_1 \geq d_2 \geq d_3 \geq ... \geq 0$, which represents the proportion of the total variation that $\xi_j(t)$ explains. Moreover, $\xi_j(t)$ are the corresponding eigenfunctions subject to $\int[\xi_j(t)]^2dt = 1$. Let $E[X(t)] = \mu(t)$, the j-th FPC score of cuvre $X_i(t)$ can be calculated as

$$\rho_{ij} = \int \xi_j(t)[X_i(t) - \mu(t)]dt.$$

However, the raw FPC functions may be rough with wiggling patterns; therefore, a smoothed FPCA is introduced to reduce the high-frequency variation in the observed $X_i(t)$ and provide better reconstruction of future $X_i(t)$ [Rice and Silverman (1991)]. The key idea of the smoothed FPCA is to impose a smoothing penalty using Sobolev norm [Adams and Fournier (2003)] to the eigenfunctions, which is

$$||\xi(t)||_L^2 = \int [\xi(t)]^2 dt + \lambda \int [L\xi(t)]^2 dt,$$

where the $[L\xi(t)]^2$ refers to the second derivative of the eigenfunctions. In other word, we aim to search the eigenfunctions $\xi(t)$ which maximizes

$$\frac{\mathrm{Var}[\int \xi_j(t) X_i(t) dt]}{\int [\xi_j(t)]^2 dt + \lambda \int [L\xi_j(t)]^2 dt}$$

Leave-one-curve-out cross-validation (CV) approach is used to choose an appropriate smoothing parameter λ, with procedure as follows [Silverman (1996)]. First, a fixed number J of FPCs is determined by the percentage of variation explained above a threshold such as 90% [Ramsay and Silverman (2005)]. Next, we remove one curve $X_i(t)$ and conduct FPCA on the remaining curves to get fitted FPC functions $\xi_1^{-i}, .., \xi_J^{-i}$. Then we see how well the FPC functions reconstruct the removed curve $X_i(t)$ by minimizing the sum of squared errors

$$R_i(\lambda) = \min \int (X_i(t) - \sum_{j=1}^{J} a_{ij} \xi_j^{-i}(t))^2 dt,$$

with respect to $a_{ij}, j = 1, \ldots, J$. Finally, the CV score for λ is $CV(\lambda) = \sum R_i(\lambda)$, and we will select the λ which minimizes the $CV(\lambda)$. In the FPCA algorithm, eigenfunctions are estimated as the linear combinations of basis functions. Using the first J FPC functions, each observed curve $X_i(t)$ can be approximated using the first J eigenfunctions as

$$X_i(t) \approx \bar{X}(t) + \sum_{j=1}^{J} \hat{\rho}_{ij} \hat{\xi}_j(t).$$

Both the FPCA and smoothed FPCA have been implemented in the function `pca.fd()` within the R package `fda` [Ramsay et al. (2009),Ramsay et al. (2018)].

2.3 Filtering methods: clustering on feature information

Filtering methods first approximate the original data into some feature information, such as the coefficients of the basis functions or FPC scores, in which the dimension of the data is reduced [James and Sugar (2003)]; then the classical clustering algorithms are conducted on the feature information as a multivariate case [Jacques and Preda (2014)]. In this paper, we applied the K-means clustering algorithm on the coefficients of the B-spline basis functions and smoothed FPC scores.

2.3.1 The classical K-means clustering

The classical K-means clustering desires to find k clusters where each object is assigned to the cluster with the nearest mean [Hartigan and Wong (1979)]. Given a set of observed multivariate data points $\{p_i\}_{i=1}^n$, the main procedure is to search K partitions $\{C_1, C_2, ..., C_k\}$ with center vectors $\{c_1, c_2, ..., c_k\}$ that minimize

$$\frac{1}{n}\sum_{j=1}^{k}\sum_{p_i \in C_j} \|p_i - c_j\|^2$$

where $\|\cdot\|$ is defined as the Euclidean norm [Hartigan and Wong (1979)]. The general schema of the K-means algorithm is

1. Initially, randomly pick k observations as the center of k clusters.
2. Assign the observations to a cluster, and update its center sequentially through alternating the following two steps until the clustering results of all objects are consistent.
   (a) Assign each observation to the cluster whose center has the least Euclidean distance to the observation,
   (b) Based on the cluster result from previous step, update means as well as the center of all k clusters.

The consistency of the K-means algorithm has been proved on the B-spline coefficients, indicating that the estimated center of clusters will converge to the unique $\{c_1^*, c_2^*, ..., c_k^*\}$ when the number of curves within each cluster increases [Abraham et al. (2003)]. In addition, the consistency of the K-means algorithm on the FPC scores has been verified through our simulation study.

2.3.2 Choosing the number of clusters

An important step in k-means algorithms is to pre-determine an appropriate number of clusters. However, it is usually an unknown and challenging problem in real applications. The general idea is to provide a set of candidate numbers and select the number with the best clustering results. In clustering analysis, a criterion to evaluate the clustering result is sometimes called an index. The R package `NbClust` includes 30 different indices to help users decide the optimal number of clusters[Charrad et al. (2014)]. For K-means clustering, 26 indices are involved in the cluster number selection, and the number voted by the majority of indices is the final number of clusters [Charrad et al. (2012)].

2.4 FunFEM: a model-based clustering method for functional data

Model-based clustering method, also called adaptive methods, for functional data introduces a mixture model that tends to identify the common patterns within and between the data [Jacques and Preda (2014)]. FunFEM is one of the model-based clustering method that introduces the discriminative functional mixture (DFM) model to estimate the probability of each observed curve belonging to a specific cluster [Bouveyron et al. (2015)]. This method was proposed for detecting the common operation patterns in the bike sharing systems (BSSs) from different European cities [Bouveyron et al. (2015)]. The algorithm has been implemented into a R package `FunFEM` with the BSSs data [Bouveyron (2015)]. In both BSSs and our U.S. weather data, the observed curves are collected from different cities or locations with their corresponding geo-information. In addition, the data were smoothed by a finite set of basis functions, resulting in some peak-like patterns. These two main similarities between BSSs and U.S. weather data motivated us to applied the FunFEM method to our data set.

2.4.1 The DFM model and the FunFEM algorithm

The main idea of FunFEM is to build up a DFM model to predict the probability of each observed curve $X_i(t)$ belonging to one of the K clusters [Bouveyron et al. (2015)]. It introduces an latent variable $Z = (Z_1, ..., Z_K) \in \{0, 1\}^K$. For each curve $X_i(t)$, $Z_k = 1$ if $X_i(t)$ belongs to k^{th} cluster, and $Z_k = 0$ otherwise. Then, for each $X_i(t)$, the probability of $Z_k = 1$ is estimated for all k using a modified expectation-maximization (EM) algorithm [Bouveyron et al. (2015)].

Remark that the n observed curves $X_i(t)$ are smoothed by the set of basis functions $B_1(t), ..., B_P(t)$ and $X_i(t) = \sum_{p=1}^{P} \beta_{ip} B_p(t)$. Let $\boldsymbol{\beta}_{P\times n} = (\beta_{ip})$ be the coefficient matrix of the basis functions. In the fitted DFM model, the smoothed curve $X_i(t)$ could be reparameterized into a new set of basis functions $\phi_1(t), ..., \phi_D(t)$ with $D < P$ and $X_i(t) = \sum_{d=1}^{D} \lambda_{id}\phi_d(t)$. Let $\boldsymbol{\Lambda}_{D\times n} = (\lambda_{id})$ be the coefficient matrix of the new basis functions [Bouveyron et al. (2015)]. Here, the new basis functions is obtained from the original basis function through a linear transformation $\phi_j(t) = \sum_{p=1}^{P} u_{jp} B_p(t)$ with the connected coefficient matrix $\boldsymbol{U}_{P\times D} = (u_{jp})$. Finally, the relationship between $\boldsymbol{\beta}, \boldsymbol{\Lambda}$ and $\boldsymbol{U}$ can be expressed as

$$\boldsymbol{\beta} = \boldsymbol{U}\boldsymbol{\Lambda} + \varepsilon,$$

where $\varepsilon \sim N(\mu_k, \Sigma_k) \in \mathbb{R}^p$ is random and independent noise. In addition, conditionally on $Z_k = 1$, $\boldsymbol{\Lambda}$ is assumed to be followed a multivariate Gaussian distribution as

$$\boldsymbol{\Lambda}_{|Z_k=1} \sim N(\mu_k, \Sigma_k),$$

where μ_k and Σ_k are the mean and the variance-covariance matrix for the k^{th} cluster. Under the above assumptions, the marginal distribution of $\boldsymbol{\beta}$ follows a mixture of Gaussian distribution as

$$P(\beta) = \sum_{k=1}^{K} \pi_k \Phi(\beta; \boldsymbol{U}\mu_k, \boldsymbol{U}\Sigma_k\boldsymbol{U}^T + E)$$

where β is the coefficients of the original basis functions $B_p(t), p = 1, ..., P$ for a curve $X_i(t)$, $\varPhi$ is the standard Gaussian density function, and $\pi_k = P(Z_k = 1)$ is the prior probability of the k^{th} cluster. The parameters of the DEM model are estimated through FunFEM, an modified EM algorithm conducted on the lower-dimension functional subspace F corresponding to D new basis functions $\phi_d(t)$ [Bouveyron et al. (2015)]. The FunFEM algorithm alternates between 3 steps at each iteration [Bouveyron et al. (2015)]:

1. The E-step computes the posterior probabilities that each curve belongs to the k^{th} cluster.
2. The F-step estimates the matrix $\boldsymbol{U}$ of the discriminative latent space conditionally on the posterior probabilities.
3. The M-step estimates the parameters of the DFM model by maximizing the conditional expectation of the complete log-likelihood.

2.4.2 Model selection and choosing the number of clusters

Three crucial parameters in FunFEM, the number of clusters, the dimension of the subspace D and the variance-covariance matrix E of noise, should be selected when conducting the algorithm. The dimension of the subspace D is selected iteratively in the algorithm. Bayesian information criterion (BIC) [Schwarz (1978)] and Integrated Completed Likelihood (ICL) Biernacki et al. (2000)] are recommended to use for selecting the number of clusters and the variance-covariance matrix E. These criteria are implemented in `FunFEM`. Given a fitted model M, the details of these criteria are as follows

- BIC(M) $= -l(\hat{\theta}) + \frac{\xi(M)}{2} log(n)$, where n is the number of observations;
- ICL = BIC - $\sum_{k=1}^{K} \sum_{i=1}^{n} Z_{ik} \times \log(z_{ik})$, where Z_{ik} is the indicator for the cluster of the i^{th} observation such that $Z_{ik} = 1$ if the i^{th} observation belongs to the k^{th} cluster and 0 otherwise.

Compared to BIC, ICL determines the number of cluster through the final allocation results of the observations; moreover, it has been observed to choose the model and cluster number with more separated cluster patterns [Schmutz et al. (2020)]. Before selection, the set of models and cluster number candidates are established respectively. Then, the values of the criterion are computed for all combinations of model and cluster number, and the one with the smallest criterion value will be selected.

3 Real Data Applications

3.1 Data description

In this section, we apply methods described in section 2 to real weather data, which comes from the 2018 Joint Statistical Meetings (JSM) Data Expo case competition and is publicly available at `https://community.amstat.org/`

`stat-computing/data-expo/data-expo-2018`. The data contains three-year forecast and historical weather records across 50 states in the United States from July 2014 to August 2017. The historical weather records comprise different weather measures in each city, such as minimum and maximum daily temperatures. The forecast weather records consist various weather measures that have been forecast over the 3-year period, and specify the date that was forecast and the date that the forecast was made on. The geographical information of the cities is also available that each city is documented with its corresponding state, geographical coordinates and airport code.

To evaluate the weather forecast performance, we defined our observations y_{ij} as the absolute value of prediction error of the minimum daily temperature in the same manner as defined in Bauer et al. (2015):

$$y_{ij} = |T_i(t_{ij})^{real} - T_i(t_{ij})^{fore}|,$$

where $T_i(t_{ij})^{real}$ and $T_i(t_{ij})^{fore}$ are the real and forecast temperatures at day t_{ij} of state i respectively, where t_{ij} is from July 2014 to August 2017. Since people are most interested in short-term weather forecasts as it provides direct guidance on planning day-to-day activities [Lazo et al. (2009)], we only evaluated the overall accuracy of one-day forecast in this study.

3.2 Results

We first reconstructed the functional form of the original data by smoothing cubic B-spline. The GCV selected 17 B-spline basis on 13 interior points, which divided the 3-year period into 14 time intervals with the same amount of data. Figure 3 demonstrates the smoothed curves of 50 states in light blue and the mean curve in red.

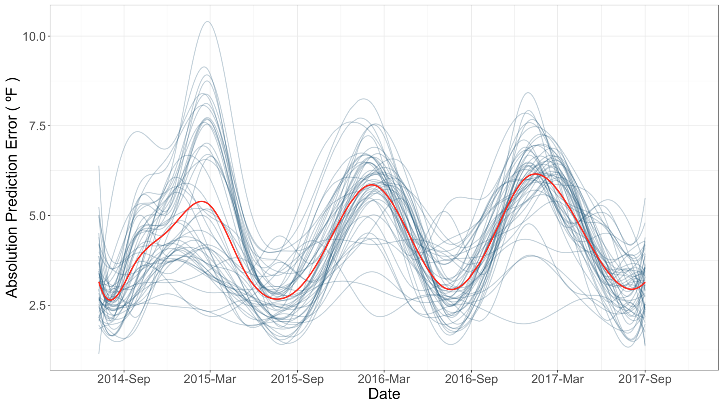


Fig. 3: Smoothed curves of the absolute prediction error for 50 U.S. states using cubic B-spline. Light blue lines represent individual curves for each state, and the red line represents the mean curve of all 50 states.

Then, the smoothed FPCA was conducted on the smoothed curves. We chose the first 5 raw FPCs based on the fact that they accounted for over 90% of total variation within the data. Leave-one-curve-out CV selected the smoothing penalty parameter; however, for the relative ease of interpretation, we selected a larger penalty parameter, which generated smoother FPC functions but the corresponding K-means clustering results were consistent. As a result, the first 5 smoothed FPCs explained 95.48% of total variation within the data.

The first 3 smoothed FPC functions are shown in Figure 4. The first FPC explains most of the variation (68.01%) and is positive over the whole time interval with local maximum at every winter time. This implies that the major source of variation is contributed by weather forecasts in winter, but such variation decreases over year. The first FPC can be interpreted as the weighted average of the absolute prediction error by 50 states over 3 years, indicating the overall variance pattern within the data. The second FPC is negative before 2016 and is positive after 2016, representing the change of the absolute prediction error over these two time intervals. The third FPC shows periodic pattern with positive values in summer and negative values in winter, so we interpreted it as the contrast between summer and winter. The fourth and fifth FPCs only explain a total of 6% variation, so they would not have large contribution to total variation, but we will still involve them in the further clustering analysis.

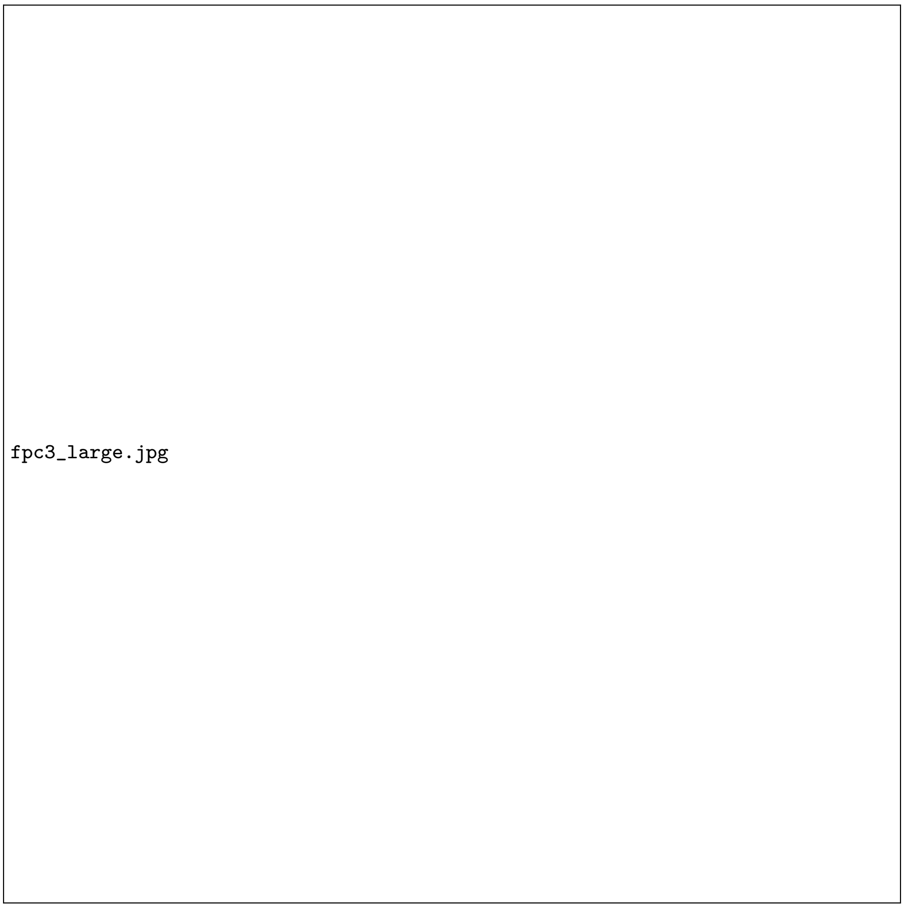
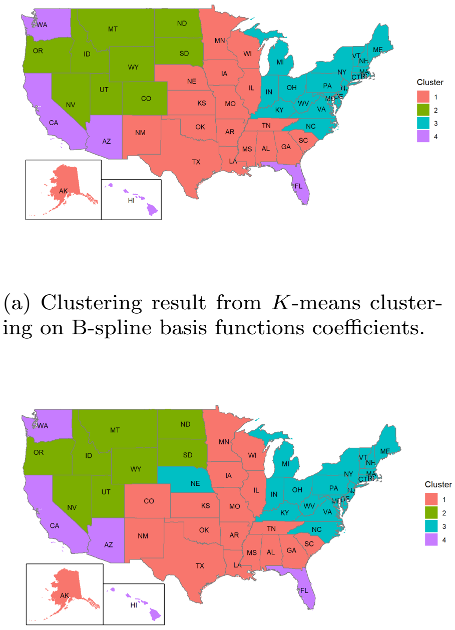


Fig. 4: The first 3 FPCs, explaining 68.01%, 11.32% and 10.32% of the variation, respectively.

The FunFEM method was used to select the number of clusters since it shows relatively better performance than the filtering methods in the simulation study in section 4. We identified four clusters by conducting the FunFEM, K-means clustering on B-spline coefficients and FPC scores. A way to assess the goodness-of-clustering is to map the states in each cluster on the actual U.S map. Figure 5 is a visualization mapping out the clustering results of the three clustering methods, where each cluster is labeled by different colour. As discussed in the earlier section, states that are close to each other or states are not neighbours but have similar climate conditions, are expected to have a similar weather prediction performance. We hope that our clustering results are able to capture this characteristic. Generally, three different clustering methods generated consistent results. Here, we identified some interesting features and similarities:

- Most states are naturally grouped together according to their spatial locations;
- All three clustering methods cluster Hawaii, Washington, California, Arizona and Florida into one cluster (cluster 4). These are the states located at the west and south coastline, which demonstrated that they share some special characteristics in weather forecast performance that makes they group together even through they are not nearby to each other;
- Most of the states in the northeast region are grouped together (cluster 3) in all three clustering methods, which indicates that the patterns of the prediction error in this region are distinctive to the ones in other areas;
- The middle inland states are divided by two clusters, where one cluster (cluster 2) includes the states in the northwest region, and the other cluster (cluster 1) includes the states in the southeast region and Alaska. Some states near the border of the cluster 1 and 2 have different assignment in different clustering methods, which are further discussed in the latter paragraphs.



(a) Clustering result from K-means clustering on B-spline basis functions coefficients.

(b) Clustering result from K-means clustering on smoothed FPC scores.

(c) Clustering result from FunFEM model-based clustering.

Fig. 5: U.S. map was partitioned into four clusters using three clustering methods.

Next, we looked into the specific patterns for each cluster. Figure 6 displays the absolute prediction error curves and the corresponding mean curves for each of the four clusters identified by the FunFEM model-based clustering. We observe that all clusters have a higher prediction error and wider confidence interval during the winter, which implies the extrapolation problem that exists in forecasting minimum temperature under cold conditions. Moreover, some special characteristics exist in each cluster:

- Cluster 1 and 2 show an increasing trend of prediction error over time, whereas cluster 1 has larger fluctuation within a year but smaller variance.
- Cluster 3 shows an annually periodic trend of prediction error over time, and its fluctuation is largest in all 4 clusters. One plausible reason might be that the states in northeast have severe winter weather (i.e. winter storm, heavy snow and cold wave) which makes the temperature more difficult to predict;

– Cluster 4 has visually lower and less fluctuated absolute prediction error than the other clusters, which means that the climate is more stable throughout the years compared to other states.

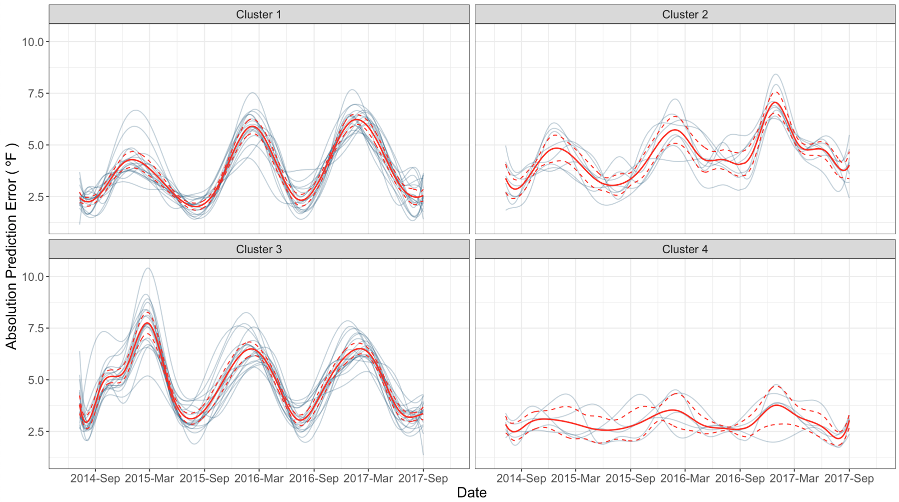


Fig. 6: Absolute prediction error curves for each of the four clusters identified by FunFEM. Light blue lines represent the individual curves for each state, whereas the red lines are the mean curves of each cluster, and red dash lines represent the 95% confidence intervals.

From Figure 5, we recognize that most of the states have a consistent result among three clustering methods, but only four of them are exception: Nebraska, Wyoming, Colorado and South Caroline. Based on their locations, they are all at the borders between the clusters. To further investigate the reason of the non-consistence, their prediction error curves and the mean curves of their assigned clusters are graphed in Figure 7. By comparing the prediction error curves of the states (black solid lines) to the mean curves of their assigned clusters in the clustering methods (colored dashed line), we observed that the curve patterns of the states prediction error were as a mixture of the patterns of the assigned clusters. For instance, the winter pattern of the curve in Wyoming is more similar to the increasing pattern of cluster 2, but the value of the curve in summer is closer to the value of the mean curve in cluster 1. The mixture pattern of the states indicate the states are in the transitional zones of different clusters.

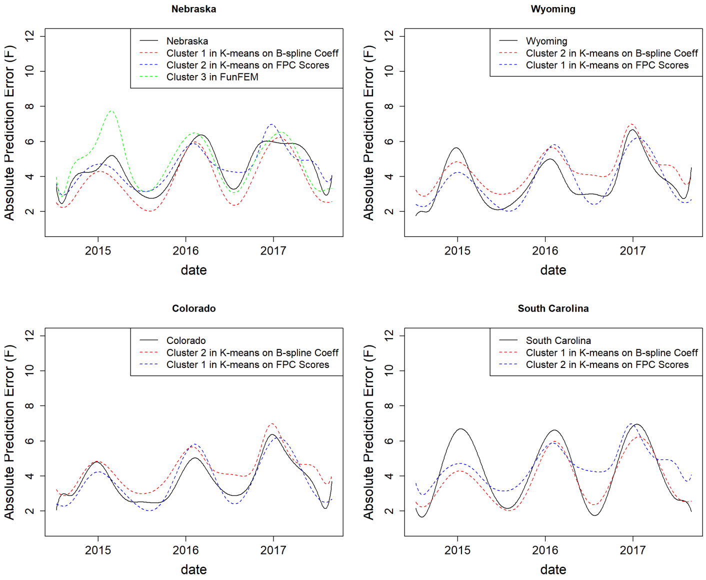


Fig. 7: Prediction error curves in Nebraska, Wyoming, Colorado and South Caroline. The black solid line is the smoothed curve of the state mentioned in title, and the dashed lines in different color are the mean curves of the clusters that include the state in different clustering methods.

Finally, we wanted to quantify cluster-to-cluster difference in prediction accuracy based on clustering results. We estimated, $\bar{y}_k$, the average daily absolute prediction error for each cluster k as

$$\bar{y}_k = \frac{1}{365 \times 3} \int_{t \in \tau} \mu_k(t) dt,$$

where the $\mu_k(t)$ was the mean curve of the cluster k. Ordering the four clusters by the magnitude of $\bar{y}_k$ for $k = 1, 2, 3, 4$, we are allowed to identify the most and least predictable states in U.S. The ranking results shown in Table 1 illustrate that cluster 4 (i.e. California, Florida, Hawaii) has the best prediction performance with the smallest $\bar{y}_k$, while cluster 3 which contains Pennsylvania, New York and Massachusetts, etc. has the worst prediction performance with the largest $\bar{y}_k$.

Rank	**Cluster**	$\bar{y}_k$	**Representative States**
1	Cluster 4	3.15	California, Florida, Hawaii
2	Cluster 1	4.05	Alaska, Texas, Louisiana
3	Cluster 2	4.68	Nevada, Oregon, North Dakota
4	Cluster 3	5.26	Pennsylvania, New York, Massachusetts

Table 1: Average daily absolute prediction error for each cluster.

4 Simulation Studies

4.1 Overall simulation setup

In this section, the performances of the three applied clustering methods were examined and compared by two aspects: the ability to select an appropriate number of clusters and to identify each curve to the correct cluster. The simulation has the following setup. First, we fixed a real number of clusters $K = 4$, and let each cluster contain an equal number of curves. Next, we defined the curves in each cluster of the following forms:

- Cluster 1: $X_i(t) = \sin(2t) + \epsilon_1(t)$
- Cluster 2: $X_i(t) = 2\sin(2t) + \epsilon_2(t)$
- Cluster 3: $X_i(t) = \frac{1}{2}\sin(t) + \epsilon_3(t)$
- Cluster 4: $X_i(t) = \sin(4t) + \epsilon_4(t)$

where the range of $t \in [0, 10]$, and $\epsilon_k(t) \sim N(0, \sigma^2)$ is the random noise for the k^{th} cluster. Then we considered two different scenarios of σ:

- Scenario 1: σ is constant for all the clusters over time.
- Scenario 2: σ is proportionally to the absolute value of the mean function, such that $\epsilon_1(t) \sim N(0, \sigma|\sin(2t)|)$, $\epsilon_2(t) \sim N(0, 2\sigma|2\sin(2t)|)$, $\epsilon_3(t) \sim N(0, \sigma|\frac{1}{2}\sin(t)|)$, and $\epsilon_4(t) \sim N(0, \sigma|\sin(4t)|)$. This scenario simulates the extrapolation case in reality.

In each scenario, each curve was generated pointwise with 1001 equidistant observed time points $t = 0, 0.01, ..., 9.99, 10$, and then was smoothed by 24 cubic B-spline basis functions with 20 intervals of equally length. We chose the number of curved n in each cluster to be 20 and let $\sigma = 2$ for each scenario. When σ is larger, we expect that the curves in the same cluster are more spread out, which is shown in Figure 8.

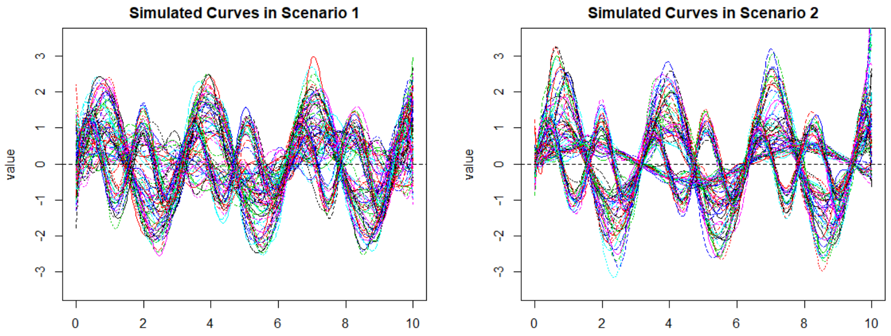


Fig. 8: Simulated curves in two different scenarios with $n = 20$ per cluster and 80 curves in total and $\sigma = 2$

4.2 Validation of cluster number selection

In this section, we assessed the ability of the filtering and model-based FunFEM method to select an appropriate number of clusters by simulation studies. For each scenario, 200 datasets have been generated through the above simulation procedures. For each generated dataset, the number of clusters were determined through 26 indices in R package `NbClust` for the K-means clustering, and BIC/ICL criterion in the R package `FunFEM` for the FunFEM method. Each clustering method suggested an optimal number of clusters from the candidate set $\{2, 3, 4, 5, 6\}$. The results are summarized in Table 2.

Scenario	Clustering Method	Selected Number of Clusters				
		2	3	4 (real)	5	6
1	K-means on B-spline coefficients	0	200	0	0	0
	K-means on FPC scores	1	85	95	18	1
	FunFEM (BIC)	0	1	**152**	44	7
	FunFEM (ICL)	0	1	**154**	31	14
2	K-means on B-spline coefficients	0	200	0	0	0
	K-means on FPC scores	2	105	82	10	1
	FunFEM (BIC)	0	0	**132**	55	13
	FunFEM (ICL)	0	0	**139**	56	5

Table 2: Frequency of the number of clusters selected over 200 simulations ($n = 20$ for each cluster) using different clustering methods in 2 different scenarios under $\sigma = 2$. The real number of clusters is 4.

Table 2 shows that the FunFEM algorithm has a dominant advantage for detecting the real number of clusters compared to the K-means. When the number of curves is small (i.e. 20 per cluster or 80 curves in total) which is similar to our case study, FunFEM reaches over 75% and 65% accuracy using BIC and ICL criterion, respectively. On the other hand, under small number of curves, the K-means clustering usually underestimates the number of clusters. The K-means clustering on B-spline coefficients underestimates the cluster number in all simulated datasets, while the K-means clustering on smoothed FPC scores underestimates about 50%. We have also investigated the methods' performance of detecting cluster number when number of curves was larger (i.e 50 curves per cluster). The underestimation problem of K-means clustering has not been improved, while the FunFEM was observed to have overestimation problem. The result of the simulation study on cluster number selection with large number of curves is attached in the supplementary documents.

4.3 Validation of clustering results

In this section, we evaluated the clustering results through different methods when the real number of clusters is given. Similarly, under different combination of scenarios (1 and 2), the number of curves within each cluster n and the variance of noise σ, we generated 200 datasets, and the real cluster labels were attached on all the curves in each generated dataset. Then, we identified four clusters using three studied clustering methods. By comparing the real labels and estimated labels obtained through the clustering procedure, the performances of the clustering methods can be evaluated by the accuracy of the clustering results. In other words, we evaluated the proportion of curves that were correctly classified into the cluster that they simulated from.

Due to the possibility of switching cluster label in clustering, before we computed the accuracy of each clustering method, we matched the label names from estimated labels to real labels through an permutation matrix Π using Hungarian algorithm [Papadimitrou and Steiglitz (1982)], which had been implemented in a function `solve_LSAP` from the R package `clue` [Hornik (2019)]. Given total N curves for clustering, the main object of this algorithm is to find the optimal permutation matrix $\Pi_{N\times N}$ so as to minimize the Euclidean partition dissimilarity [Dudoit and Fridlyand (2002)]. Using the Π, we switch the labels of estimated clusters to match labels in real cluster; then the accuracy is computed as the proportion of the curves labels that are matched between the real and estimated clustering result.

Scenario	Methods	Mean Accuracy (SD)
1	K-means on B-spline Coefficients	0.852 (0.179)
	K-means on FPC Scores	0.839 (0.175)
	FunFEM with BIC	**0.987 (0.065)**
	FunFEM with ICL	**0.983 (0.072)**
2	K-means on B-spline Coefficients	0.793 (0.180)
	K-means on FPC Scores	0.816 (0.174)
	FunFEM with BIC	**0.976 (0.080)**
	FunFEM with ICL	**0.987 (0.062)**

Table 3: Summary table of average accuracy from the clustering results over 200 simulations in 2 different scenarios under $\sigma = 2$. The real number of clusters is 4 and there are $n = 20$ curves per cluster.

Table 3 summarizes the clustering results from simulated data. The best clustering methods suggested by Table 3 is FunFEM, which outperforms other two methods with over 95% accuracy of clustering the correct curves to the true cluster. We also have computed two more metrics, the average within-cluster standard deviation (AWSD) and the average distance of cluster-specific mean function between estimated and real clusters (ADCMF). A better clustering result is expected to have a lower value of them as they indicate a more reasonable clustering result with better estimation on real within-cluster mean curve and less within-cluster variation. FunFEM also performs the best among these 3 clustering methods using AWSD and ADCM. Further details of AWSD and ADCMF results are shown in Table 5 and 6 in the supplementary documents.

4.4 Additional simulation

Based on the real data in Figure 9 below, we observe 4 clusters with spatial characteristics. We conduct another simulation based on the temporal and spatial patterns. The new simulation setup is given below:

1. Treat our clustering results (results of FunFEM) as the true clusters. For each cluster k, we perform FPCA and get eigenvalue λ_{jk} and eigenfunction $\xi_{jk}(t)$. We chose the first 3 FPCs which explained over 90% of the total variation within the data. Let us denote $i = 1, ..., 20$ as the index of curves, $j = 1, 2, 3$ as the index of FPCs, $k = 1, 2, 3, 4$ as the index of clusters.
2. For each cluster k, we simulate 20 curves with the following procedures:
   (a) Simulate FPC scores $\hat{\rho}_{ijk} \sim N(0, \lambda_{jk})$.
   (b) Reconstruct the ith curve of the kth cluster as

$$\tilde{X}_{ik}(t) = \mu_k(t) + \sum_{j=1}^{3} \hat{\rho}_{ijk}\xi_{jk}(t). \tag{1}$$

3. Perform 5 clustering methods discussed in the paper (K-means on raw data, B-spline coefficients, FPC scores and FunFEM with BIC and ICL criteria) on simulated curves.

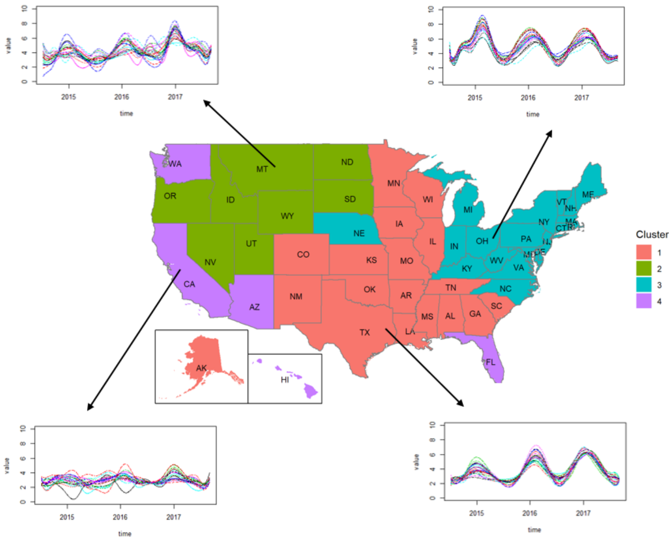


Fig. 9: U.S. map of 4 clusters obtained from FunFEM algorithm and simulated curves correspond to each cluster. Each cluster includes 20 simulated curves.

Scenario	Clustering Method	Selected Number of Clusters				
		2	3	4 (real)	5	6
Real data simulation	K-means on Raw Data	200	0	0	0	0
	K-means on B-spline coefficients	10	110	59	17	4
	K-means on FPC scores	5	178	16	1	0
	FunFEM (BIC)	0	0	**105**	21	74
	FunFEM (ICL)	0	0	**112**	23	65

Table 4: Frequency of the number of clusters selected over 200 simulations ($n = 20$ for each cluster) using different clustering methods on simulated curves corresponding to the 4 clusters obtained from FunFEM algorithm. The true number of clusters is 4.

Scenario	Methods	Mean Accuracy (SE)
Real data simulation	K-means on Raw Data	0.863 (0.00895)
	K-means on B-spline Coefficients	0.846 (0.0068)
	K-means on FPC Scores	0.851 (0.0078)
	FunFEM with BIC	**0.893 (0.0061)**
	FunFEM with ICL	**0.894 (0.0053)**

Table 5: Summary table of average accuracy from the clustering results over 200 simulations on simulated curves corresponding to the clusters obtained from FunFEM algorithm. The real number of clusters $K = 4$ is assumed known.

According to Table 4 and Table 5, we observe that the FunFEM model-based clustering method outperform the traditional K-means clustering method. For selecting the correct number of cluster, FunFEM achieves over 50% of accuracy to obtain the correct number of clusters while the other K-means methods result in an accuracy less than 30%. Especially for the K-means clustering on raw data, the real number of clusters is always underestimated over 200 simulations. Therefore, the model-based functional clustering method, FunFEM, is more capable to capture the real number of clusters. In addition, given the real number of clusters, the overall accuracy of the clustering results from FunFEM is slightly higher than the ones from K-means with a lower standard error. The results of this simulation study imply that the FunFEM model-based clustering method has better performance compared to the traditional K-means clustering method in simulated data which reflect both temporal and spatial characteristics of U.S. temperature data.

5 Conclusion

This work is motivated by analyzing the weather forecast data obtained from 2018 JSM. FPCA and functional data clustering were proposed to analyze the weather prediction error over a three-year period, where we explored the major modes of variability among these 50 states over time and discovered some distinct patterns from the four clusters. We also compared three different functional data clustering methods, FunFEM, K-means clustering on B-spline coefficients and FPC scores. The majority curves in a given cluster are consistent across different methods. This provides us with more confidence with our clustering results. It is also evident that there exists spatial relationship in prediction errors, for example neighbouring states are grouped into the same cluster. States are far away but with a similar climate such as California, Hawaii and Florida are also seen in the same cluster. The most predictable cluster is cluster 4, where the daily prediction error is 3.15 Fahrenheit on average. Cluster 4 includes states, such as California and Florida, where the weather is hot and maintains a similar temperature all year round. The least predictable cluster is cluster 3, which is off by 5.26 Fahrenheit daily on average. Cluster 3 includes states, such as New York and Massachusetts, where they share a hot weather in the summer and cold weather in the winter.

Our simulation study compares the three clustering methods used in the real data application. When the true number of clusters is 4, our simulation results show that the FunFEM model-based clustering method can detect the number of clusters more accurately, as well as obtaining a higher clustering accuracy, in comparison with the other two methods, K-means clustering on B-spline coefficients and K-means clustering on FPC scores. On the other hand, when the true number of cluster is 1, we observe that both K-means clustering on B-spline coefficients and K-means clustering on FPC scores can detect the number of clusters more accurately whereas FunFEM tends to overestimate the number of clusters. In reality, the true number of clusters is usually unknown, then one suggested solution for cluster detection is conducting both K-means and FunFEM, and visualizing each cluster with their mean curves and confidence bands. Visualizing the clusters' mean curves can help check whether there are potential outliers or underlying subgroup patterns. Comparing the overlap area of the confidence bands between two clusters can provide information for the identification of the number of clusters.

Acknowledgements The authors are most appreciative of the organizers of 2018 JSM Data Expo who made this happen. The authors would also like to thank the guest editor and two reviewers for the constructive comments, which are very helpful to improve this work. The authors also thank Dr. Peijun Sang, Yuping Yang and Zhiyang Zhou, faculty members and graduate students in the Department of Statistics and Actuarial Science at Simon Fraser University who provided helpful suggestions relating to this project.

References

Abraham C, Cornillon PA, Matzner-Løber E, Molinari N (2003) Unsupervised curve clustering using b-splines. Scandinavian journal of statistics 30(3):581–595

Adams RA, Fournier JJ (2003) Sobolev spaces, vol 140. Elsevier, Atlanta

Adams RM, Rosenzweig C, Peart RM, Ritchie JT, McCarl BA, Glyer JD, Curry RB, Jones JW, Boote KJ, Allen Jr LH (1990) Global climate change and us agriculture. Nature 345(6272):219–224

Adelfio G, Chiodi M, D'Alessandro A, Luzio D (2011) FPCA algorithm for waveform clustering. Journal of Communication and Computer 8(6):494–502

Bauer P, Thorpe A, Brunet G (2015) The quiet revolution of numerical weather prediction. Nature 525(7567):47–55

Besse PC, Cardot H, Stephenson DB (2000) Autoregressive forecasting of some functional climatic variations. Scandinavian Journal of Statistics 27(4):673–687

Biernacki C, Celeux G, Govaert G (2000) Assessing a mixture model for clustering with the integrated completed likelihood. IEEE transactions on pattern analysis and machine intelligence 22(7):719–725

Bosq D (1996) Nonparametric statistics for stochastic processes: estimation and prediction, vol 110. Springer-Verlag, New York

Bouveyron C (2015) funFEM: Clustering in the Discriminative Functional Subspace. URL `https://CRAN.R-project.org/package=funFEM`, r package version 1.1

Bouveyron C, Côme E, Jacques J (2015) The discriminative functional mixture model for a comparative analysis of bike sharing systems. The Annals of Applied Statistics 9(4):1726–1760

Box GE, Jenkins GM, Reinsel GC, Ljung GM (2015) Time series analysis: forecasting and control, 5th edn. John Wiley & Sons, Hoboken, New Jersey

Cao J, Ramsay J (2009) Generalized profiling estimation for global and adaptive penalized spline smoothing. Computational statistics & data analysis 53(7):2550–2562

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2012) NbClust package: finding the relevant number of clusters in a dataset. UseR! 2012

Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust: An R package for determining the relevant number of clusters in a data set. Journal of Statistical Software 61(6):1–36

Collomb G (1983) From non parametric regression to non parametric prediction: Survey of the mean square error and original results on the predictogram. In: Specifying Statistical Models, Springer, pp 182–204

Curry HB, Schoenberg IJ (1966) On Pólya frequency functions IV: the fundamental spline functions and their limits. Journal d'analyse mathématique 17(1):71–107

Dudoit S, Fridlyand J (2002) A prediction-based resampling method for estimating the number of clusters in a dataset. Genome biology 3(7):1–21

Györfi L, Härdle W, Sarda P, Vieu P (1989) Nonparametric curve estimation from time series, vol 60. Springer-Verlag, New York

Hartigan JA, Wong MA (1979) Algorithm as 136: A k-means clustering algorithm. Journal of the Royal Statistical Society Series C (Applied Statistics) 28(1):100–108

Hornik K (2019) clue: Cluster ensembles. URL `https://CRAN.R-project.org/package=clue`, r package version 0.3-57

Jacques J, Preda C (2014) Functional data clustering: a survey. Advances in Data Analysis and Classification 8(3):231–255

James GM, Sugar CA (2003) Clustering for sparsely sampled functional data. Journal of the American Statistical Association 98(462):397–408

Ke Y, Li J, Zhang W, et al. (2016) Structure identification in panel data analysis. The Annals of Statistics 44(3):1193–1233

Lazo JK, Morss RE, Demuth JL (2009) 300 billion served: Sources, perceptions, uses, and values of weather forecasts. Bulletin of the American Meteorological Society 90(6):785–798

Li J, Yue M, Zhang W (2019) Subgroup identification via homogeneity pursuit for dense longitudinal/spatial data. Statistics in Medicine 38(17):3256–3271

Lin Z, Wang L, Cao J (2016) Interpretable functional principal component analysis. Biometrics 72(3):846–854

Luo W, Cao J, Gallagher M, Wiles J (2013) Estimating the intensity of ward admission and its effect on emergency department access block. Statistics in medicine 32(15):2681–2694

Orrell D, Smith L, Barkmeijer J, Palmer T (2001) Model error in weather forecasting. Nonlinear processes in geophysics 8(6):357–371

Papadimitrou CH, Steiglitz K (1982) Combinatorial optimization: algorithms and complexity. Prentice-Hall, New York

Radhika Y, Shashi M (2009) Atmospheric temperature prediction using support vector machines. International Journal of Computer Theory and Engineering 1(1):55–59

Ramsay J, Silverman B (2005) Functional data analysis, 2nd edn. Springer, New York

Ramsay J, Hooker G, Graves S (2009) Functional data analysis with R and MATLAB. Springer, New York

Ramsay JO, Wickham H, Graves S, Hooker G (2018) fda: Functional Data Analysis. URL `https://CRAN.R-project.org/package=fda`, r package version 2.4.8

Rice JA, Silverman BW (1991) Estimating the mean and covariance structure nonparametrically when the data are curves. Journal of the Royal Statistical Society: Series B (Methodological) 53(1):233–243

Sang P, Wang L, Cao J (2017) Parametric functional principal component analysis. Biometrics 73(3):802–810

Schmutz A, Jacques J, Bouveyron C, Cheze L, Martin P (2020) Clustering multivariate functional data in group-specific functional subspaces. Computational Statistics DOI ff10.1007/s00180-020-00958-4ff

Schwarz G (1978) Estimating the dimension of a model. The Annals of Statistics 6(2):461–464

Silverman BW (1996) Smoothed functional principal components analysis by choice of norm. The Annals of Statistics 24(1):1–24