

# Functional data clustering using K-means and random projection with applications to climatological data

Mehdi Ashkartizabi<sup>1</sup> · Mina Aminghafari<sup>1</sup>

Published online: 7 August 2017  
© Springer-Verlag GmbH Germany 2017

**Abstract** In this paper, an efficient pattern recognition method for functional data is introduced. The proposed method works based on reproducing kernel Hilbert space (RKHS), random projection and K-means algorithm. First, the infinite dimensional data are projected onto RKHS, then they are projected iteratively onto some spaces with increasing dimension via random projection. K-means algorithm is applied to the projected data, and its solution is used to start K-means on the projected data in the next spaces. We implement the proposed algorithm on some simulated and climatological datasets and compare the obtained results with those achieved by K-means clustering using a single random projection and classical K-means. The proposed algorithm presents better results based on mean square distance (MSD) and Rand index as we have expected. Furthermore, a new kernel based on a wavelet function is used that gives a suitable reconstruction of curves, and the results are satisfactory.

**Keywords** Functional data · K-means algorithm · Random projections · Reproducing kernel Hilbert space · Climatological data

## 1 Introduction

On the one hand, the goal of clustering is to discover homogeneous groups in the data. Clustering approaches are divided into hierarchical and non-hierarchical methods which K-means is in the second group, (Everitt et al. 2011, Chapters 4 and 7).

On the other hand, functional data appear in many fields such as physical processes, genetics, biology, meteorology and signal processing (Bernardi et al. 2017; Ferraty and Vieu 2006; Ruiz-Medina and Espejo 2012; Bagirov and Mardaneh 2006). “A random variable  $\mathbf{X}$  is called a functional variable if it takes values in an infinite dimensional space (or a functional space). An observation  $X$  of  $\mathbf{X}$  is a functional data” (Ferraty and Vieu 2006, Chapter 1), and its analysis is called functional data analysis (FDA). FDA is an important topic in statistics and has received a wide range of applications. It contains principal components analysis, canonical correlation and discriminant analysis, functional linear models and clustering. Ramsay and Silverman (2005) presented various techniques to analyze these data.

Generic examples of functional data contain weather data and human growth data. Some authors such as Balzanella et al. (2017), Bohorquez et al. (2017) and Ruiz-Medina (2014) used FDA for environmental science application. Mateu and Romano (2017) presented a comprehensive reviewed of spatial functional data issue. In medicine, the growth of a child and the evaluation state of a patient are essentially functional data (Jacques and Preda 2014b; Tokushige et al. 2007) although their observations are discrete. They are primarily functional data because they can be considered as a continuous function of another variable such as time.

✉ Mina Aminghafari  
aminghafari@aut.ac.ir

Mehdi Ashkartizabi  
mehdi90@aut.ac.ir

<sup>1</sup> Department of Statistics, Faculty of Mathematics and Computer Science, Amirkabir University of Technology, Tehran, Iran

In practice, functional data are observed from a finite set of time-points. In other words, we usually have a set of  $\{f_i(t_1), \dots, f_i(t_n)\}$  with  $i = 1, \dots, N$ ,  $t_j \in [0, T]$  and  $T \in \mathbb{R}^+$ . The matrix representation of these observations is as follows:

$$\begin{bmatrix} f_1(t_1) & f_1(t_2) & \cdots & f_1(t_n) \\ f_2(t_1) & f_2(t_2) & \cdots & f_2(t_n) \\ \vdots & \vdots & \ddots & \vdots \\ f_N(t_1) & f_N(t_2) & \cdots & f_N(t_n) \end{bmatrix}.$$

Each row of this matrix corresponds to one observation such as growth of a child which is recorded at  $n$  points. If  $n$  tends to infinity, then each row becomes a continuous curve (i.e. infinite dimensional data or functional data). Raw data do not include extra information involved by an underlying process which produces data. Due to continuous nature of data, if a multivariate analysis is used on the raw data, the suitable results may not be achieved. In functional data clustering, the suitable result is a performance of an algorithm for assigning data to true clusters. It can be measured with different indices such as mean square distance. There are three shortcomings of multivariate analysis (Ramsay 1982) for these types of data: (1) more investigation about variations e.g. estimation of the first and second derivative along with the time, is impossible, (2) realization dependency in separate points of time is ignored, (3) information about variable between two consequence time-point is not available. Hence, FDA is appropriate. To use FDA including functional data clustering, it is common to reconstruct a continuous curve from finite points. It is done by approximating curves through some finite basis functions such as splines (specially B-splines), Fourier series, wavelets and reproducing kernel Hilbert space (RKHS). The last one is used in this paper. To this end, let  $\mathcal{H}$  be a real Hilbert space of functions and a set  $\{\phi_1, \dots, \phi_L\}$ ,  $L \in \mathbb{N}$  be an orthonormal basis that generated some space of functions in  $\mathcal{H}$ . Each  $\phi_l$ ,  $l = 1, \dots, L$ , is a continuous function, for example in Fourier series  $\phi_l(t) = A_l \cos(lt) + B_l \sin(lt)$  where  $A_l$  and  $B_l$  are constant. Then, a function  $f_i \in \mathcal{H}$  can be expressed as a linear combination as follows:

$$f_i(t) = \sum_{l=1}^L \alpha_{il} \phi_l(t) = \Phi^T(t) \alpha_i, \quad (1)$$

where  $\Phi(t) = (\phi_1(t), \dots, \phi_L(t))^T$  and the expansion coefficient  $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iL})^T$  of  $i$ -th curve,  $f_i$ , is estimated using an interpolation procedure or least squares smoothing (Jacques and Preda 2014a). The continuous estimation of  $f_i$  is  $\hat{f}_i(t) = \Phi^T(t) \hat{\alpha}_i$ .

Functional data are infinite dimensional data because they are continuous curves. In the most of functional data clustering methods, the expansion coefficients,  $\alpha_i$ , obtained

from Eq. (1) are clustered, instead of clustering the functional data, see e.g. James and Sugar (2003), Abraham et al. (2003) and Luz-López-García et al. (2015). In other words, they project the infinite dimensional data onto the finite basis functions  $\{\phi_1, \dots, \phi_L\}$  and obtain  $\alpha_i$ . Then, they cluster the projected data,  $\alpha_i$ , which are finite dimensional data. For clustering  $\alpha_i$ , K-means algorithm is used commonly.

In this paper, we focus on functional data clustering using RKHS and random projection.

In the next paragraph, we review some papers in functional data clustering. In these articles, functional data are projected onto a basis function and using K-means  $\alpha_i$  are clustered. The main purpose of these works is to improve K-means algorithm. The common basis applied in these works is spline basis. Although they have suitable properties such as easy usage, and mathematical description (James and Sugar 2003; Jacques and Preda 2014a), they have two main drawbacks: (1) they applied to a function with some degrees of regularity and they are not appropriate for an irregular function and a function with peaks, (2) they require heavy computational for high dimensional data (Antoniadis et al. 2013; Giacofci et al. 2013). For these reasons, some works used wavelet decomposition which is suitable for irregular function and non-stationary data (Antoniadis et al. 2013).

Abraham et al. (2003) projected data onto a B-spline basis and clustered the expansion coefficients via K-means algorithm using Euclidean distance. Yamamoto (2012) proposed a new objective function for this aim. Giuseppe et al. (2013) presented a method for clustering Italian climate zones using B-spline. Antoniadis et al. (2013) considered K-means algorithm clustering and proposed a dissimilarity measure between curves based on wavelet transform. Ieva et al. (2013) clustered the multivariate functional data via K-means algorithm using a proper distance. Luz-López-García et al. (2015) projected the original data onto RKHS and reduced the dimension of coefficients by two strategies: functional principal components and random projection. Then, they clustered the coefficients by K-means algorithm. Abramowicz et al. (2017) presented a non-parametric statistical method which gives a suitable result to cluster misaligned dependent curves for seasonal climate interpretation.

In the following, we present various works which increased the performance of functional data clustering approaches such as non-parametric approach and model-based approach. Also, some of them compare different methods. James and Sugar (2003) projected the functional data onto spline basis and proposed the model-based algorithm that is especially for sparse data. Rossi et al. (2004) projected data onto spline basis and used self organized map (SOM) algorithm to cluster expansion

coefficients. Ray and Mallick (2006) decomposed the data with wavelet transform and clustered obtained coefficients using a model-based algorithm. Bouveyron and Jacques (2011) applied a model-based method on the functional principal components. Boullé (2012) proposed non-parametric density estimation to cluster functional data. Haggarty et al. (2012) developed univariate and multivariate functional data clustering to determine a pattern of water quality data in Scotland. Jacques and Preda (2013) applied the model-based clustering via probability density which is defined by Delaigle and Hall (2010). The model-based clustering is extended to the multivariate case by Jacques and Preda (2014b). Jacques and Preda (2014a) presented a classification of different methods for functional data clustering: (1) raw-data clustering, (2) two-stage methods, (3) non-parametric clustering, (4) model-based clustering. Finazzi et al. (2015) presented a novel approach for functional data clustering and compared it with other approaches such as K-means algorithm.

By a regionalization procedure, groups of sites which are homogenous in term of hydro-meteorological characteristics could be found. One of the useful techniques to divide catchments in a region into natural groups is clustering. In what follows, we present several works which are related to regionalization of climatological data. Some of those works e.g. Rao and Srinivas (2006b) and Srinivas (2008) are based on hybrid clustering and global fuzzy K-means.

Mitchell (1976) regionalized the western United States into climatic regions based on the air masses. Anyadike (1987) used cluster analysis to the regions over West Africa, and obtained ten-region climatic clusters. Brring (1988) regionalized the daily rainfall in Kenya using common factor analysis. El-Jabi et al. (1998) divided the province of New Brunswick (Canada) into homogeneous regions by the regionalization of floods. Alila (1999) regionalized the precipitation annual maxima in Canada using a hierarchical approach. Guenni and Bárdossy (2002) disaggregated the highly seasonal monthly rainfall data using a stochastic method in two steps. Rao and Srinivas (2006a) regionalized the annual maximum flow data from the watersheds in Indiana by fuzzy cluster analysis. The regionalization of watersheds by hybrid-cluster analysis (the hybrid of Wards and K-means algorithms) was proposed by Rao and Srinivas (2006b). They conclude that the hybrid method is flexible and presents suitable results for regionalization studies. Srinivas (2008) combines self-organizing feature map and fuzzy clustering to regionalization of watersheds. Bharath and Srinivas (2015b) delineated extreme rainfall in India by global fuzzy c-means based on location indicators (latitude, longitude, and altitude) and seasonality measure of extreme rainfall. Asong et al. (2015) proposed a new approach for

regionalization of precipitation climate characteristics in the Canadian prairie provinces. The obtained regions are homogeneous statistically and climatologically. Nam et al. (2015) delineated the climatic rainfall regions of South Korea using a multivariate analysis and regional rainfall frequency analysis. Rahman et al. (2017) developed the regional flood frequency analysis techniques by generalized additive models for Australia. The proposed method obtains better results than three other methods (log-linear model, canonical correlation analysis and region-of-influence approach).

Most of the mentioned methods are based on K-means algorithm (James and Sugar 2003; Ieva et al. 2013; Antoniadis et al. 2013; Luz-López-García et al. 2015). The K-means algorithm starts by initializing the  $k$  cluster centers randomly and minimizing the total distance (total square error) between each of the  $N$  points and its nearest centroid. The cluster centroid is the arithmetic mean of the points on each cluster. The distance is minimized when the algorithm converges, however, it may be a local minimum. In general, the objective function of K-means is not convex, and there is no guarantee that its solution is global. It reaches to a local solution in some datasets, (Cardoso and Wichert 2012).

Before analyzing high dimensional data, their dimension usually is reduced. The principal component analysis (PCA) (Jolliffe 2002) and random projection (RP) (Johnson and Lindenstrauss 1984) are used for this purpose. In this paper and some other works (Luz-López-García et al. 2015; Cardoso and Wichert 2012), high dimensional data are projected onto a lower dimensional subspace via RP such that Euclidean distances between the points are approximately preserved.

In this paper, we propose a new algorithm which has the following advantages:

1. It improves all known K-means based functional data clustering approaches.
2. It avoids local minimum which leads to a solution (i.e. a set of labels for curves) with a lower mean square distance (MSD) and larger Rand index.
3. A new kernel based on a wavelet function reduces the condition number of a linear system and MSD. In most of the data, MSD, Rand index and the number of iterations of the proposed method based on the wavelet kernel are smaller than other ones.
4. K-means in the last step of the proposed method, has less iteration number than clustering using a single random projection and classical K-means.
5. Since the obtained condition number under the new kernel is small, it produces a suitable and stable reconstruction of the curves that may be used in the estimation (prediction) problem of FDA.

The paper is organized as follows: Sect. 2 briefly presents RKHS, random projection and random projection K-means algorithm for functional data clustering denoted by (RPFDF). The proposed method is introduced in Sect. 3. Section 4 reports the results of simulation studies. The real datasets are clustered in Sect. 5. Conclusions are given in Sect. 6.

## 2 RKHS and RPFDF K-means

In this section, we present a brief review of RKHS and applying of Tikhonov regularization theory in RKHS. For the general version of Tikhonov regularization and theory of RKHS see (Tikhonov and Arsenin 1997; Aroszajn 1950), respectively. An excellent background for RKHS representation is given by González-Hernández (2010). At the end of this section, we mention the introduced algorithm by Luz-López-García et al. (2015) which called RPFDF.

### 2.1 RKHS

**Definition 1** (Reproducing kernel) Let  $\mathcal{H}$  be a real Hilbert space of functions defined in  $X \subset \mathbb{R}^n$  with inner product  $\langle \cdot, \cdot \rangle_{\mathcal{H}}$  and the norm  $\|f\|_{\mathcal{H}} = \langle f, f \rangle_{\mathcal{H}}^{1/2}$ . A function  $K : X \times X \rightarrow \mathbb{R}$  is called a reproducing kernel of  $\mathcal{H}$  if

1.  $K(\cdot, x) \in \mathcal{H}$  for all  $x \in X$ ,
2.  $f(x) = \langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}}$  for all  $f \in \mathcal{H}$  and for all  $x \in X$ .

A Hilbert space of functions is called a reproducing kernel Hilbert space (RKHS) if it admits a reproducing kernel. From Definition 1, it is concluded that  $K(x, y) = \langle K(\cdot, x), K(\cdot, y) \rangle_{\mathcal{H}}$  and for each  $f \in \mathcal{H}$  of the form  $f(x) = \sum_{i=1}^n \alpha_i K(x, x_i)$  with  $x, x_i \in X$  and  $\alpha_i \in \mathbb{R}$ , we have

$$\|f\|_{\mathcal{H}}^2 = \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) = \boldsymbol{\alpha}^T \mathbf{K}_x \boldsymbol{\alpha}. \quad (2)$$

Now, we express the Tikhonov regularization theory to reconstruct functions from sample points. Let compact subset  $X \subset \mathbb{R}^p$ ,  $K$  be a Mercer kernel (positive definite and symmetric kernel),  $\mathcal{H}_K$  is its associated RKHS and the Borel probability measure in  $X \times \mathbb{R}$  is denoted by  $\nu$ . Let  $f_\nu(x) = \int_{\mathbb{R}} y d_\nu(y|x)$  be a regression function where  $d_\nu(y|x)$  is the conditional probability measure on  $\mathbb{R}$ . Both  $\nu$  and  $f_\nu$  are unknown, and we want to reconstruct  $f_\nu$  using a random sample. Let  $X_n := \{x_1, \dots, x_n\} \subset X$  and  $f^j := \{(x_i, y_{ij}) \in X \times \mathbb{R}\}_{i=1}^n$  be a random sample with  $j = 1, \dots, N$ . Let  $V_n := \text{span}\{K(\cdot, x_i) : x_i \in X_n\}$  be a function space (the span of an arbitrary set may be defined as the set of all finite linear combinations of its elements).

Tikhonov regularization theory projects the sample  $f_n$  onto  $V_n$ . It builds a stable reconstruction of  $f_\nu$  by solving the

$$f_n^{*j} := \arg \min_{f^j \in V_n} \frac{1}{n} \sum_{i=1}^n (f^j(x_i) - y_{ij})^2 + \gamma \|f^j\|_{\mathcal{H}_K}^2 \quad (3)$$

where  $\gamma > 0$  is a regularization parameter,  $j = 1, \dots, N$  and  $\|f\|_{\mathcal{H}_K} = \langle f, f \rangle_{\mathcal{H}_K}^{1/2}$ . In other words, Tikhonov regularization theory states that there is a stable solution for the optimization problem (3). The following theorem gives a representation form for  $f_n^{*j} \in V_n \subset \mathcal{H}_K$ .

**Theorem 1** (Luz-López-García et al. 2015) *Under the mentioned assumptions there is a unique solution of (3) which can be represented by*

$$f_n^{*j}(x) = \sum_{i=1}^n \alpha_i K(x, x_i), \quad \text{for all } x \in X$$

and  $j = 1, \dots, N$ ,

where  $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^T$  is obtained from

$$(n\gamma \mathbf{I}_n + \mathbf{K}_x) \boldsymbol{\alpha} = \mathbf{y}_j, \quad (4)$$

where  $\mathbf{I}_n$  is an identity  $n \times n$  matrix,  $\mathbf{y}_j = (y_{1j}, \dots, y_{nj})^T$  and matrix  $(\mathbf{K}_x)_{st} = K(x_s, x_t)$  for  $s, t = 1, \dots, n$ . The estimate of  $f_\nu$  in  $X_n$  is as follows:

$$\hat{f}_n^{*j} = \mathbf{K}_x \boldsymbol{\alpha}. \quad (5)$$

Since  $\mathbf{K}_x$  is a Mercer kernel, it is decomposed as  $\mathbf{K}_x = V^T D V$ , where the rows of  $V$  are eigenvectors of the  $\mathbf{K}_x$  and the elements of diagonal matrix  $D$  are corresponding (non-negative) eigenvalues. Therefore, we can write

$$\mathbf{K}_x = (D^{1/2} V)^T (D^{1/2} V) = U^T U, \quad \text{where } U = D^{1/2} V. \quad (6)$$

Using Eqs. (6) and (2),  $\|f\|_{V_n}^2 = \|U\boldsymbol{\alpha}\|^2$  which means that the norm of  $f \in V_n$  is equal to the Euclidean norm of its transformation. Let  $\nu$  be a Borel probability measure defined on  $X \times \mathbb{R}$ . Suppose that  $f_n^*(x) = \sum_{i=1}^n \alpha_i K(x, x_i)$  and  $g_n^*(x) = \sum_{i=1}^n \beta_i K(x, x_i)$  are the regularized  $\gamma$ -projection of  $f_\nu$  and  $g_\mu$  respectively, which obtained from two samples  $f_n := \{(x_i, y_i) \in X \times \mathbb{R}\}_{i=1}^n$  and  $g_n := \{(x_i, y'_i) \in X \times \mathbb{R}\}_{i=1}^n$ . It follows by Luz-López-García et al. (2015) that the distance between  $f_n$  and  $g_n$  is attained by

$$d_{V_n}^2(f_n, g_n) := \|f_n^* - g_n^*\|_{V_n}^2 = (\boldsymbol{\alpha} - \boldsymbol{\beta})^T \mathbf{K}_x (\boldsymbol{\alpha} - \boldsymbol{\beta}) = \|U\boldsymbol{\alpha} - U\boldsymbol{\beta}\|^2.$$

### 2.2 Random projection and RPFDF

The nature of functional data requires high CPU times. So, we need to reduce their dimension. This fact

encourages us to use RP for dimensionality reduction. The RP projects  $N$  points of dimension  $n$  to a random subspace of dimension  $h$ . To use RP, we generate a random matrix that its elements are randomly drawn from any zero mean distribution. For example, let

$$A = \begin{bmatrix} 2 & 3 & 2 & 0 & 4 \\ -2 & 3 & 12 & 10 & 9 \\ 0 & -3 & 4 & 1 & 0 \\ 4 & 5 & -1 & 7 & -4 \end{bmatrix} \text{ be data in } \mathbb{R}^5$$

(four points of dimension 5). To project them onto  $\mathbb{R}^2$ ,

$$\text{we generate a random matrix } R = \begin{bmatrix} 2.63 & 1.52 \\ 0.44 & 0.50 \\ -0.50 & 1.75 \\ -0.15 & 1.17 \\ -0.53 & -0.16 \end{bmatrix}$$

which elements are drawn from the standard normal distribution. The projected data in  $\mathbb{R}^2$  are obtained as follows:

$$B = A \times R = \begin{bmatrix} 3.49 & 7.38 \\ -16.15 & 29.67 \\ -3.46 & 6.67 \\ 14.26 & 15.65 \end{bmatrix} \text{ (four points of dimension 2).}$$

The success of RP is highlighted by the Johnson-Lindenstrauss lemma (Johnson and Lindenstrauss 1984; Dasgupta and Gupta 2003) is presented as follows:

**Lemma 1** For any real number of  $\varepsilon, \delta \in (0, 1)$  and any  $N \in \mathbb{N}$ , let  $h \in \mathbb{N}$  such that

$$h \geq 4 \left( \frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3} \right)^{-1} \log \left( \frac{N}{\sqrt{\delta}} \right)$$

then, with probability  $1 - \delta$

$$(1 - \varepsilon) \|U\alpha - U\beta\| \leq \|RU\alpha - RU\beta\| \leq (1 + \varepsilon) \|U\alpha - U\beta\|,$$

where  $R$  is a  $h \times n$  matrix and  $\alpha, \beta$  are obtained from Eq. (4) and  $U$  is attained from Eq. (6).

Lemma 1 states that the Euclidean distances between the points and the Euclidean distances between their projections are approximately preserved. The distance between  $f_n$  and  $g_n$  is defined as  $d_{v_n}(f_n, g_n) = \|U\alpha - U\beta\|$ . According to Lemma 1,  $d_{v_n}(f_n, g_n) = \|U\alpha - U\beta\| \cong \|RU\alpha - RU\beta\|$  and the projection matrix is  $P = RU$ . Now, we can cluster the reduced expansion coefficients,  $\alpha$ . A review of RPFDF is presented in Algorithm 1.

**Algorithm 1** Random projection K-means for functional data (RPFDF)

**input:**

1. Dimension  $h$  (dimension of the projected space).
2. The number of clusters  $k$ .
3.  $X_n := \{x_1, \dots, x_n\} \subset X$  and a sample of empirical functions  $\{f^j\}_{j=1}^N$  where  $f^j := \{(x_i, y_{ij}) \in X \times \mathbb{R}\}_{i=1}^n$ .

**output:**

1. Cluster label  $G$ .

**begin**

1. Calculate the kernel matrix  $K_x$  by  $(K_x)_{ij} = K(x_i, x_j)$ .
2. Compute the matrix  $U = D^{\frac{1}{2}}V$  where  $K_x = (D^{\frac{1}{2}}V)^T(D^{\frac{1}{2}}V) = U^T U$ .
3. Choose the regularization parameter  $\gamma$  based on cross-validation.
4. For  $j = 1, \dots, N$ , solve  $(\gamma n \mathbf{I}_n + K_x)\alpha = y^j$  where  $y^j = (y_{1j}, \dots, y_{nj})^T$ . Denote the solution by  $\alpha^j$ .
5. Initialize  $G$  randomly.
6. Generate the random matrix  $R_{h \times n}$ .
7. Set the projection matrix  $P_{h \times n} = R_{h \times n} U$ .
8. For  $j = 1, \dots, N$ , project the data using  $\beta_{h \times 1}^{jT} = P_{h \times n} \alpha_{n \times 1}^j$  and put them in the rows of a matrix  $\mathcal{A}_{N \times h}^{RP}$ .
9. According to  $G$ , calculate the mean of each cluster in  $\mathcal{A}_{N \times h}^{RP}$  and put them in rows of a matrix  $C_{k \times h}^{RP}$  as centers of the clusters.
10. Apply K-means on  $\mathcal{A}_{N \times h}^{RP}$  with  $C_{k \times h}^{RP}$  as initialization to obtain  $G$ .
11. Return  $G$ .

**end**

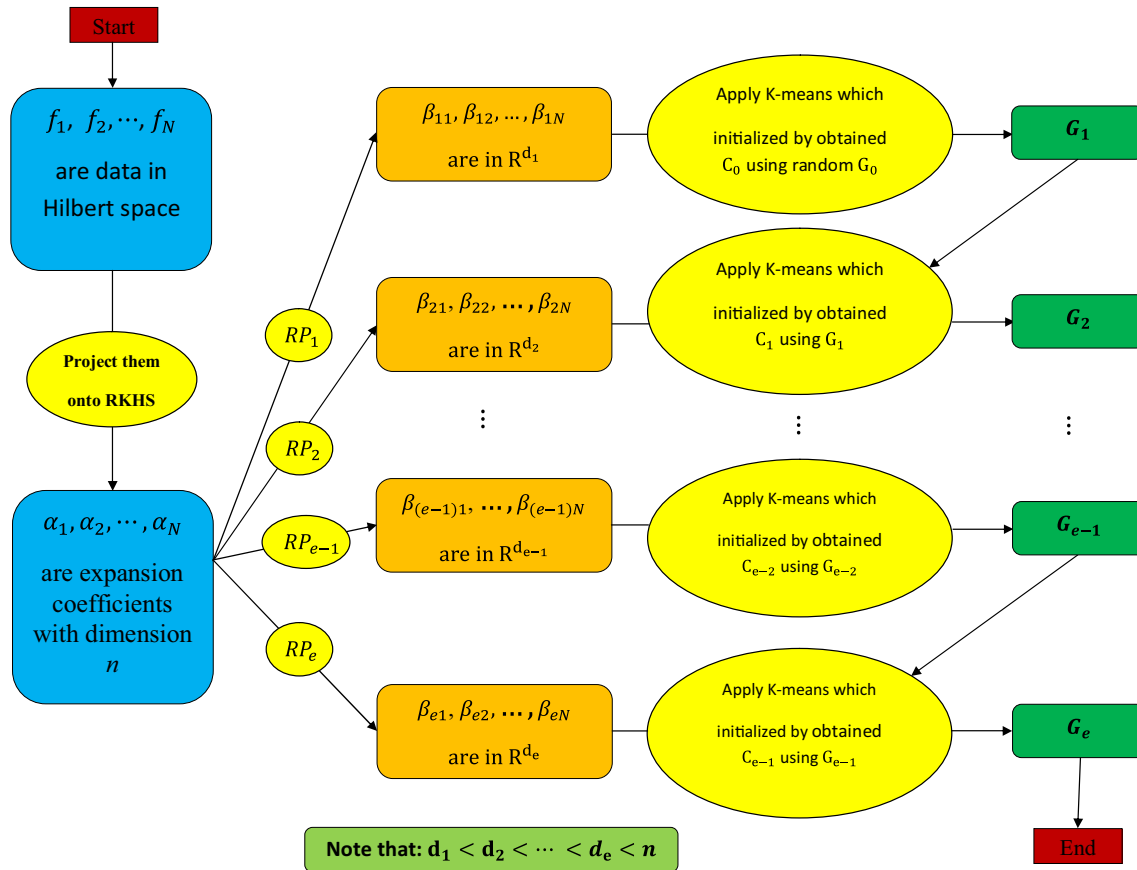
### 3 Proposed method

In Sect. 2, we reviewed an algorithm for functional data clustering. It first projects the curves onto RKHS and second RP reduces the dimension of expansion coefficients. Then, it clusters them using K-means algorithm. There are two difficulties in this algorithm: (1) projecting data onto RKHS and reducing the dimension by RP cause more details are lost and the probability of assigning a point to a wrong cluster increases. (2) clustering by K-means algorithm may give a local solution (i.e. a set of labels).

To overcome these difficulties, we propose an algorithm to improve the Algorithm 1. We call the proposed algorithm iterative RP functional data clustering (IRPFDF). The scheme of this algorithm is presented in Fig. 1. Note that when we use iterative RP, K-means attains a non-local solution. We explain IRPFDF in what follows.

As we mentioned before, RPFDF uses a RP onto a single space, but in IRPFDF, we project the expansion coefficients onto several spaces then we use K-means to cluster them. In each of the epochs, we increase the dimension of the projection. In a given dimension, the obtained clusters are used to initialize the K-means in the next dimension.





**Fig. 1** The flowchart of the proposed method.  $G$  is a set of cluster labels

Let  $X_n := \{x_1, \dots, x_n\} \subset X$  and a sample of empirical functions  $\{f^j = (x_i, y_{ij}), i = 1, \dots, n\}_{j=1}^N$  be data. To project data onto RKHS, we calculate the kernel matrix  $K_x$  which is the size of  $n \times n$ . We choose a suitable regularization parameter  $\gamma$  using cross-validation (CV which will be explained briefly for a real dataset). The expansion coefficients  $(\alpha_{n \times 1}^j)$  are obtained from Eq. (4) for all curves. We also compute Cholesky decomposition of  $K_x$  from Eq. (6) to obtain  $U_{n \times n}$ . After projecting data onto RKHS (i.e. obtaining expansion coefficients  $\alpha_{n \times 1}^j$ ), we project the expansion coefficients onto a random subspace of dimension  $d_1 < n$  as follows. For dimension  $d_1$ , we generate the random matrix  $R_{d_1 \times n}$  and obtain the projection matrix  $P_{d_1 \times n}$  from  $P_{d_1 \times n} = R_{d_1 \times n} U_{n \times n}$ . Now, we project  $\alpha_{n \times 1}^j$  using  $\beta_{d_1 \times 1}^j = P_{d_1 \times n} \alpha_{n \times 1}^j$  and put  $\beta_{d_1 \times 1}^j$  in the rows of a matrix  $\mathcal{A}_{N \times d_1}^{RP_1}$ . To initialize label of clusters, we generate  $N$  random numbers from  $unif(\{1, \dots, k\})$  where  $unif$  is discrete uniform distribution and put them in the set  $G$ . Using  $G$ , we obtain the cluster centers,  $C_{k \times d_1}^{RP_1}$ , based on the mean of each cluster in  $\mathcal{A}_{N \times d_1}^{RP_1}$ . By applying K-means on  $\mathcal{A}_{N \times d_1}^{RP_1}$  which is initialized by  $C_{k \times d_1}^{RP_1}$ ,  $G$  is updated. Based on  $G$ , we obtain

$C_{k \times d_2}^{RP_2}$  from  $\mathcal{A}_{N \times d_2}^{RP_2}$ . We initialize K-means for dimension  $d_2$  using  $C_{k \times d_2}^{RP_2}$  to cluster matrix  $\mathcal{A}_{N \times d_2}^{RP_2}$ . We repeat this process for some subspaces with increasing dimension i.e.  $d_1 \leq d_2 \leq \dots \leq d_e < n$ . The relation between  $d_z$  for  $z = 1, \dots, e$  is heuristic, and we can find them using cross-validation.

Lower dimension is like a high temperature in simulated annealing clustering, and when we increase the dimension, it cools the simulated annealing clustering. So, the relation between  $d_z$  is similar to simulated annealing cooling (Cardoso and Wichert 2012; Selim and Alsultan 1991). The details of the proposed method are given in Algorithm 2.

**Algorithm 2** Iterative random projection K-means for functional data (IRPFDD)

**input:**

1. List of dimension  $d_z, z = 1, \dots, e$  (dimension of the projected spaces).
2. The number of clusters  $k$ .
3.  $X_n := \{x_1, \dots, x_n\} \subset X$  and a sample of empirical functions  $\{f^j\}_{j=1}^N$  where  $f^j := \{(x_i, y_{ij}) \in X \times \mathbb{R}\}_{i=1}^n$ .

**output:**

1. Cluster label  $G$ .

**begin**

1. Calculate the kernel matrix  $K_x$  by  $(K_x)_{ij} = K(x_i, x_j)$ .
2. Compute the matrix  $U = D^{\frac{1}{2}}V$  where  $K_x = (D^{\frac{1}{2}}V)^T(D^{\frac{1}{2}}V) = U^T U$ .
3. Choose the regularization parameter  $\gamma$  based on cross-validation.
4. For  $j = 1, \dots, N$ , solve  $(\gamma n \mathbf{I}_n + K_x)\alpha = y^j$  where  $y^j = (y_{1j}, \dots, y_{nj})^T$ . Denote the solution by  $\alpha^j$ .
5. Initialize  $G$  randomly.  
FOR  $z = 1$ : e
- L1. Generate the random matrix  $R_{d_z \times n}$ .
- L2. Set the projection matrix  $P_{d_z \times n} = R_{d_z \times n} U$ .
- L3. For  $j = 1, \dots, N$ , project the data using  $\beta_{d_z \times 1}^{jT} = P_{d_z \times n} \alpha_{n \times 1}^j$  and put them in rows of a matrix  $\mathcal{A}_{N \times d_z}^{RP_z}$ .
- L4. According to  $G$ , calculate the mean of each cluster in  $\mathcal{A}_{N \times d_z}^{RP_z}$  and put them in the rows of a matrix  $C_{k \times d_z}^{RP_z}$  as centers of the clusters.
- L5. Apply K-means on  $\mathcal{A}_{N \times d_z}^{RP_z}$  with  $C_{k \times d_z}^{RP_z}$  as initialization to obtain  $G$ .

end

6. Return  $G$ .

**end**

**4 Simulation studies**

In this section, we implement three methods: classical K-means, RPFDF K-means and IRPFDF K-means (the proposed method) under three kernels used by Luz-López-García et al. (2015) and a new kernel, on some simulated and real datasets that have been used in the most papers.

**4.1 Simulated data**

We describe following datasets which their figures are presented at the end of the paper.

1. **Waves** (Luz-López-García et al. 2015): For these data, there are  $N = 450$  realizations of curves in  $n = 400$  time-points, and they have  $k = 3$  clusters. Let
 
$$h_1(x) := \max(6 - |x - 11|, 0); \quad h_2(x) = h_1(x - 4);$$

$$h_3(x) = h_1(x + 4),$$

where  $x \in [1, 21]$ . The waves data are obtained by a convex combination as follows:

$$y_1(x) = uh_1(x) + (1 - u)h_2(x) + \varepsilon(x) \quad \text{for class 1,}$$

$$y_2(x) = uh_1(x) + (1 - u)h_3(x) + \varepsilon(x) \quad \text{for class 2,}$$

$$y_3(x) = uh_2(x) + (1 - u)h_3(x) + \varepsilon(x) \quad \text{for class 3,}$$

where  $u$  is a uniform random variable in  $(0, 1)$  and  $\varepsilon(x)$  is a standard normal variable. The first row of Fig. 2a corresponds to these data and their centroids.

2. **Irrationals** (Luz-López-García et al. 2015): For these data, there are  $N = 300$  realizations of curves in  $n = 1009$  time-points, and they generated from  $k = 3$  clusters. These data are not equidistant. Three irrational functions constructed as follows:

$$y_1(x) = x(1 - x) + \varepsilon(x) \quad \text{for class 1,}$$

$$y_2(x) = x^{\frac{3}{2}}(1 - x) + \varepsilon(x) \quad \text{for class 2,}$$

$$y_3(x) = x^2(1 - x) + \varepsilon(x) \quad \text{for class 3,}$$

where  $\varepsilon(x)$  is a uniform random variable in  $(-0.2, 0.2)$ . The last row of Fig. 2a shows them.

3. **Three groups**: (Ferreira and Hitchcock 2009) This dataset contains three main groups of signal functions. In this paper, we call them by group 1, group 2 and group 3. In each group, there are  $k = 4$  clusters. For each group, there are  $N = 120$  realizations of curves in  $n = 200$  time-points. They are shown in Fig. 2b.

**4.2 Kernels**

We use three kernels applied by Luz-López-García et al. (2015),

and a new kernel, Mexican hat wavelet. We calculate their parameters using cross-validation (CV) technique. We will explain it briefly for a real dataset. These kernels are as follows:

1. Laplacian kernel:

$$K(x_i, x_j) := e^{-\frac{1}{\sigma_1} \|x_i - x_j\|_1}$$

2. Gaussian kernel:

$$K(x_i, x_j) := e^{-\frac{1}{\sigma_2} \|x_i - x_j\|_2^2}$$

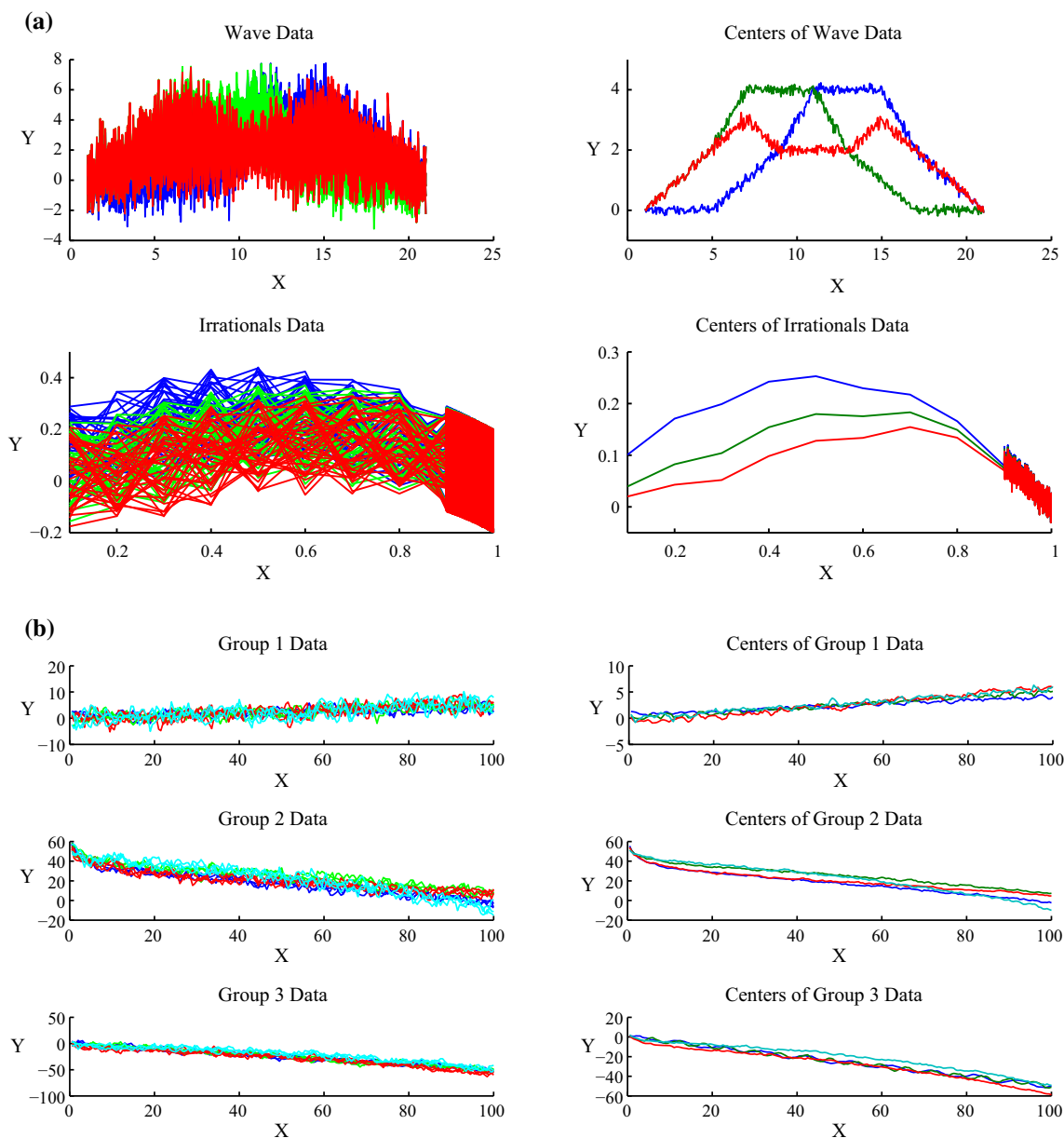
3. Polynomial kernel:

$$K(x_i, x_j) := (1 + ax_i^T \cdot x_j)^b$$

4. Mexican hat kernel:

$$K(x_i, x_j) := (\sigma_3^2 - \|x_i - x_j\|_2^2) e^{-\frac{1}{\sigma_3} \|x_i - x_j\|_2^2}$$

where  $\sigma_1, \sigma_2, \sigma_3, a > 0, b \in \mathbb{N}$  and  $\|\cdot\|_p$  is norm  $p$ .



**Fig. 2** Simulated datasets and their centroids. **a** Waves and irrationals datasets and their centroids, **b** three groups data and their centroids

Hereafter, we call the fourth kernel by *wavelet*. The kernel parameters and  $\gamma$  for simulated data are given in Table 1. Note that in *three groups* dataset, the presented parameters are the same for groups 1, 2 and 3 except for group 3,  $\sigma_3 = 0.7$ .

### 4.3 Experiments

We measure the performance of three methods (classical K-means, RPFK-means and IRPFK-means) on the reported datasets with MSD and Rand index. We also calculate running time and the number of iterations that

K-means converges. We compare the condition number of four kernels for all datasets. Now, we explain them.

The MSD is the mean square distance between each cluster center and its members. In other words, MSD is obtained by

$$MSD = \frac{1}{N} \sum_{j=1}^k \sum_{f \in S_j} \|f - C_j\|_2^2 \tag{7}$$

where  $S_j$  is the  $j$ -th cluster and  $C_j$  is its center.

Rand index (Rand 1971; Hubert and Arabie 1985) measures the similarity between two partitions. It takes a



**Table 1** Kernel parameters of simulated datasets

Data	Kernels			
	Laplacian $\sigma_1, \gamma_1$	Gaussian $\sigma_2, \gamma_2$	Polynomial a, b, $\gamma_3$	Wavelet $\sigma_3, \gamma_4$
Waves	$1, 10^{-4}$	$1, 10^{-4}$	1, 5, 1	$0.3, 10^{-4}$
Irrationals	$10, 10^{-4}$	$5, 10^{-4}$	$10^{-2}, 5, 10^{-4}$	$1, 10^{-4}$
Three groups	$1, 10^{-4}$	$1, 10^{-4}$	$10^{-4}, 5, 10^{-4}$	$0.76, 10^{-4}$ except for group 3, $\sigma_3 = 0.7$

value between 0 and 1. When it tends to 1, two partitions have a high agreement.

It can be calculated using the function `rand.index()` of the package `fossil` of the R software.

To obtain expansion coefficients, we solve Eq. (4) by numerical algorithms commonly. The type of matrix  $K_x$  has an effective role on the accuracy of solution of Eq. (4) i.e.  $\alpha$ . In the following, we consider one validity criterion which shows an accuracy of the linear system. An important issue in a numerical analysis is the condition number (Trefethen and Bau 1997) of a matrix that is associated with the linear equation  $Ax = b$ . It states the amount of sensitivity of a system. The higher condition number is equivalent to more sensitivity and yields the ill-posed system. In other words, a small variation of  $b$  makes big changes in  $x$ . The condition number of a nonsingular  $A_{n \times n}$  is defined by  $\text{cond}(A) = \|A\|_2 \cdot \|A^{-1}\|_2$ . In general, if the condition number of a matrix  $A$  be  $10^m$ , then  $m$  digits of accuracy may lose. In next sections, for each kernel and dataset, we calculate the condition number of kernel matrix,  $K_x$ , and  $T = (n\gamma I_n + K_x)$  defined in (4). We implement mentioned algorithms on the simulated and real datasets, separately.

### 4.3.1 Performance on simulated datasets

We generate data from stated models. We run each algorithm 1000 times for each dataset. In all experiments, the entries of the random matrix are independent and identically distributed standard Gaussian random variables,  $N(0, 1)$ . The classical K-means algorithm uses raw data, i.e. it uses only  $y_{ij}$  for  $i = 1, \dots, n$  and  $j = 1, \dots, N$ . The way of choosing  $h$  and  $d_z$  in the RPF and IRPF, respectively, is heuristic. In other words, we run the algorithms for different values and choose the best one based on MSD and Rand index. In Table 2, we present the average running time (by `time`), the average number of iterations, (by `Iter`), the average of MSD (by `MSD`), the average of Rand index (by `RI`) and their standard deviation by `SD` on 1000 iterations. The bolded numbers are the best results, and the selected values of  $h$  and  $d_z$  are mentioned in the table. The values are rounded at most to 2-decimal places. The cells with 0.00 have an exact value less than

0.005 which we round them up to 0.00. According to Table 2, we conclude that

1. IRPF K-means has the minimum MSD among RPF K-means and the classical K-means. In other words, it avoids local minimum;
2. As expected, in all datasets, the running time of the classical K-means is less than others because it performs K-means only one time on raw data. This is while the RPF and IRPF project the raw data onto some spaces and this needs more time than the classical K-means. The running time of IRPF K-means and RPF K-means are approximately the same in the most datasets. Note that K-means on the projected data takes a shorter running time than on the raw data;
3. The Iter of IRPF K-means is less than two other methods in all datasets except for irrationals dataset;
4. The Rand index of IRPF K-means is the best among all methods in all datasets;
5. In the most datasets for both cases (IRPF K-means and RPF K-means), the MSD, the Iter and Rand index of the wavelet kernel are better than three other kernels.

Thus, in all datasets, clustering via functional nature (especially IRPF) is better than classical K-means based on MSD, RI and Iter.

The results of the condition number are shown in Table 3. In the three groups, the results for group 1, group 2 and group 3 are equal. In all datasets, the wavelet kernel has the best condition number for the matrix  $T$ , that are bolded.

For the dataset of group 1, the boxplots of its results are presented in Fig. 3. In this figure, the classical method is denoted by `C`, RPF using wavelet kernel is denoted by `Wrp`, RPF using Laplacian kernel is denoted by `Lrp`, RPF using the Gaussian kernel is denoted by `Nrp`, RPF using the polynomial kernel is denoted by `Prp`, IRPF using wavelet kernel is denoted by `Wirp`, IRPF using Laplacian kernel is denoted by `Lirp`, IRPF using the Gaussian kernel is denoted by `Nirp` and IRPF using the polynomial kernel is denoted by `Pirp`. According to these boxplots, the wavelet IRPF has the best performance with

**Table 2** The clustering results of simulated data: Time, Iter, MSD and RI for different kernels. SD is a standard deviation

Datasets		Classical	RPFDP				IRPFDP			
			Laplacian	Gaussian	polynomial	wavelet	Laplacian	Gaussian	polynomial	wavelet
Waves	Iter	1.65	14.98	14.83	11.57	5.95	13.57	13.38	<b>0.43</b>	0.53
	SD	0.87	5.73	5.55	6.14	2.74	5.58	5.34	0.34	0.65
$h = 90$	Time (s)	<b>0.01</b>	0.25	0.21	0.66	0.49	0.26	0.22	0.66	0.50
	SD	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$d_z = (10, 20, 60, 90)^T$	MSD	713	953	955	948	689	953	955	948	<b>686</b>
	SD	46.58	11.87	11.08	8.89	17.08	11.76	10.75	8.71	3.10
	RI	0.93	0.54	0.54	0.57	0.97	0.56	0.55	0.55	<b>0.99</b>
	SD	0.17	0.07	0.02	0.01	0.04	0.00	0.00	0.00	0.01
Irrationals	Iter	<b>1.10</b>	3.83	3.77	3.84	3.29	4.24	4.23	4.06	3.64
	SD	0.48	2.06	1.96	2.06	1.80	2.17	2.19	2.12	1.95
$h = 100$	Time (s)	<b>0.07</b>	1.65	1.40	4.16	3.00	1.67	1.41	14.18	3.01
	SD	0.00	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
$d_z = (10, 20, 60, 100)^T$	MSD	13.04	13.05	13.04	13.05	13.05	<b>13.03</b>	<b>13.03</b>	<b>13.03</b>	<b>13.03</b>
	SD	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
	RI	0.53	0.52	0.52	0.52	0.52	<b>0.55</b>	<b>0.55</b>	<b>0.55</b>	<b>0.55</b>
	SD	0.02	0.03	0.03	0.03	0.03	0.00	0.00	0.00	0.00
Group 1	Iter	4.88	4.11	3.93	5.69	5.42	4.02	4.35	4.57	<b>3.47</b>
	SD	2.74	2.58	2.41	2.68	3.02	1.92	2.05	2.16	1.86
$h = 80$	Time (s)	<b>0.00</b>	0.13	0.11	0.36	0.24	0.14	0.12	0.37	0.25
	SD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$d_z = (10, 20, 50, 80)^T$	MSD	527	540	542	533	528	525	530	529	<b>515</b>
	SD	13.18	11.69	11.06	8.74	11.55	8.72	8.80	8.62	8.24
	RI	0.57	0.49	0.48	0.56	0.57	0.66	0.64	0.63	<b>0.70</b>
	SD	0.12	0.10	0.09	0.06	0.09	0.02	0.02	0.01	0.02
Group 2	Iter	1.96	4.32	4.93	5.68	3.03	1.10	1.68	5.06	<b>0.20</b>
	SD	1.42	2.35	2.54	2.57	1.85	1.32	1.55	2.34	0.59
$h = 60$	Time (s)	<b>0.00</b>	0.13	0.11	0.36	0.23	0.13	0.16	0.37	0.24
	SD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$d_z = (5, 20, 40, 60)^T$	MSD	2.34	2.39	2.44	4.06	2.33	2.28	2.30	4.04	<b>2.26</b>
	SD	( $\times 10^3$ ) 225	215	208	238	202	96	99	241	115
	RI	0.93	0.88	0.85	0.57	0.93	0.93	0.91	0.63	<b>0.96</b>
	SD	0.08	0.06	0.05	0.04	0.07	0.06	0.05	0.01	0.06
Group 3	Iter	2.03	5.09	5.46	6.39	3.33	1.18	2.24	3.50	<b>0.43</b>
	SD	1.59	2.82	2.91	2.93	2.21	1.36	1.72	1.93	0.88
$h = 60$	Time (s)	<b>0.00</b>	0.13	0.11	0.37	0.24	0.13	0.12	0.38	0.24
	SD	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$d_z = (5, 20, 40, 60)^T$	MSD	1.86	1.89	1.93	2.99	1.86	<b>1.81</b>	1.84	2.97	<b>1.81</b>
	SD	( $\times 10^3$ ) 170	131	123	226	155	48	46	207	64
	RI	0.90	0.85	0.80	0.71	0.90	0.93	0.87	0.78	<b>0.94</b>
	SD	0.19	0.08	0.07	0.05	0.10	0.06	0.04	0.03	0.07

respect to MSD and RI. We also show the scatter plot of the first and the second principal components of the true partition of  $y_1, \dots, y_N$  (top-left) and those obtained by

classical method (top-right), IRPFDP (bottom-left) and RPFDP via wavelet kernel (bottom-right) for group 1 and group 2 datasets in Fig. 4a, b, respectively. The clusters

**Table 3** The condition number of simulated data and  $T = (n\gamma\mathbf{I}_n + K_x)$  from Eq. (4)

Datasets		Laplacian	Gaussian	polynomial	wavelet
Waves	cond( $K_x$ )	$1.57 \times 10^3$	$3.70 \times 10^{19}$	$1.30 \times 10^{26}$	$9.60 \times 10^{18}$
	cond( $T$ )	604	882	$1.64 \times 10^{12}$	<b>15.5</b>
Irrationals	cond( $K_x$ )	$2.01 \times 10^9$	$2.92 \times 10^{20}$	$1.56 \times 10^{20}$	$4.07 \times 10^{20}$
	cond( $T$ )	$1 \times 10^4$	$1 \times 10^4$	$7.04 \times 10^{13}$	<b><math>9.92 \times 10^3</math></b>
Three groups	cond( $K_x$ )	16.65	$9.62 \times 10^3$	$7.90 \times 10^{19}$	<b>14.61</b>
	cond( $T$ )	15.47	174.98	$6.24 \times 10^4$	<b>11.39</b>

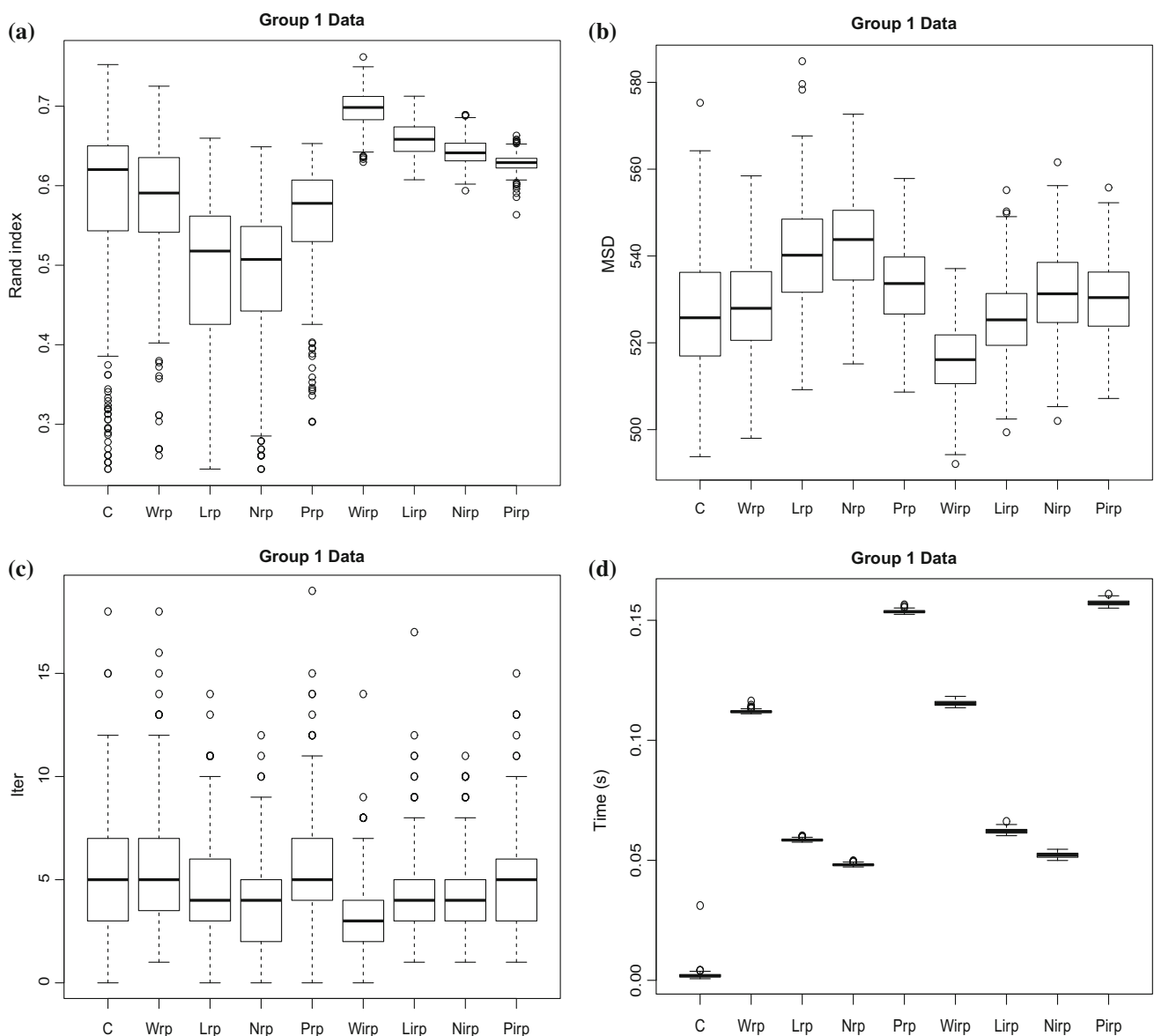
of group 1 dataset are close together and their recognitions are difficult, but for group 2 dataset, the clusters are distinct. The IRPFD method presents a partition which is more similar than others to the true partition.

### 5 Application to real datasets

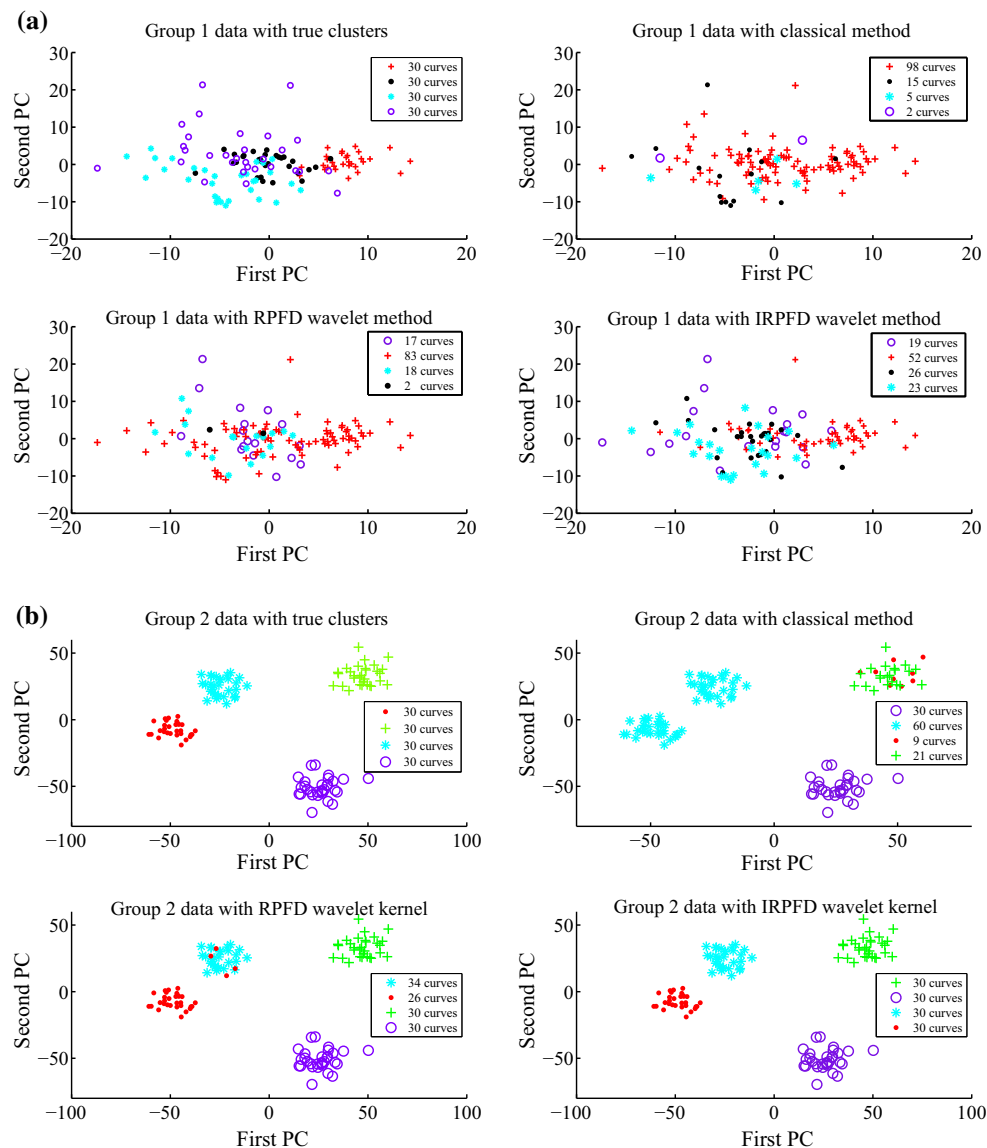
In this part, we try to obtain homogeneous groups (clusters) for real datasets. We introduce real datasets and explain CV, determining the number of clusters (using Wilks Lambda) and IRPFD for one dataset.

#### 5.1 Real data

Some examples of functional data appear in meteorology and earth science (Tokushige et al. 2007; Luz-López-García et al. 2015; Ferraty and Vieu 2006). We describe



**Fig. 3** Boxplot of Rand index, MSD, Iter and running time for group 1 data



**Fig. 4** The first and the second principal components of the different methods for group 1 and group 2 datasets

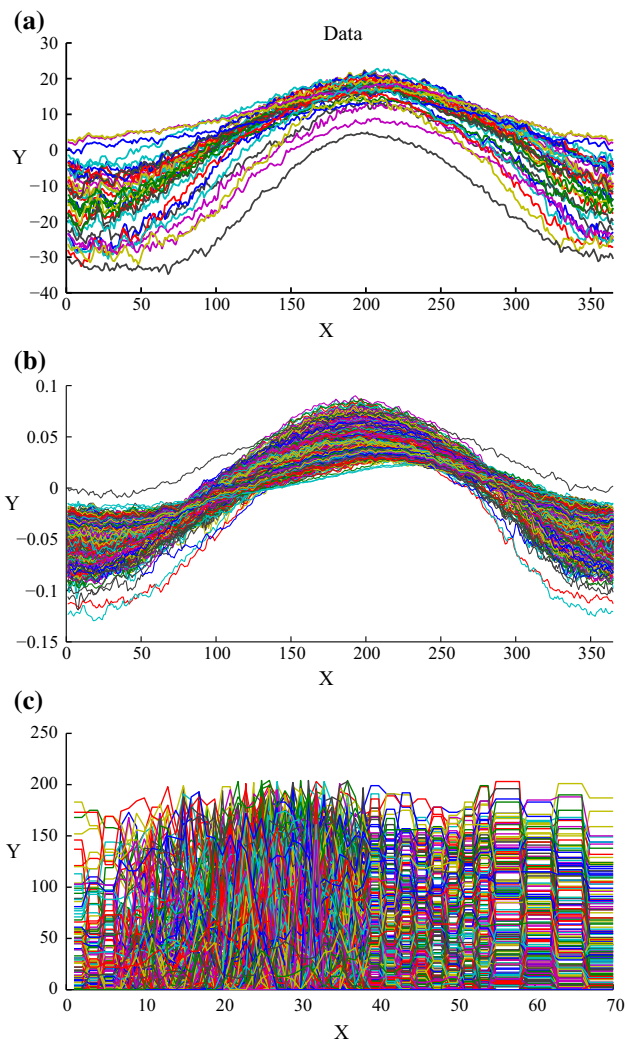
three following datasets to show the performance of the proposed method. Their figures are presented at the end of the paper.

1. Canadian temperature: Ramsay and Silverman (2005) presented the details of these data. They are available in the R package `fda`. This dataset is used by some authors for FDA and clustering, for example, see Aguilera-Morillo et al. (2017) and Jacques and Preda (2014b). For  $N = 35$  locations in Canada, the daily temperature had been measured and averaged over 1960–1994. The obtained data constitute a  $35 \times 365$  matrix whose elements correspond to the average of 35 observed temperatures over 35 years. We consider each of the rows of this matrix as a curve. We present the data in Fig. 5 and their geographical locations are

shown in Fig. 6a. Our purpose is to present the homogeneous groups of this dataset.

2. Russian temperature:<sup>1</sup> These data investigated by Bulygina and Razuvaev (2012) and are presented in detail in CDIA (Carbon dioxide information center). The source of this dataset contains temperature of 518 stations in Russia. These data have been recorded since 1874–2011. We consider 490 stations which cover an interval 1978–2009, i.e. 32 years. For each station, after removing its trend (if it exists, for example, the stations 76 and 77 are stationary), the average of temperature over 32 years is computed. Therefore, we have a matrix of size  $490 \times 365$  such that each row corresponds to the average of temperature over 32

<sup>1</sup> This dataset is available at [http://cdiac.ornl.gov/ftp/russia\\_daily/](http://cdiac.ornl.gov/ftp/russia_daily/).

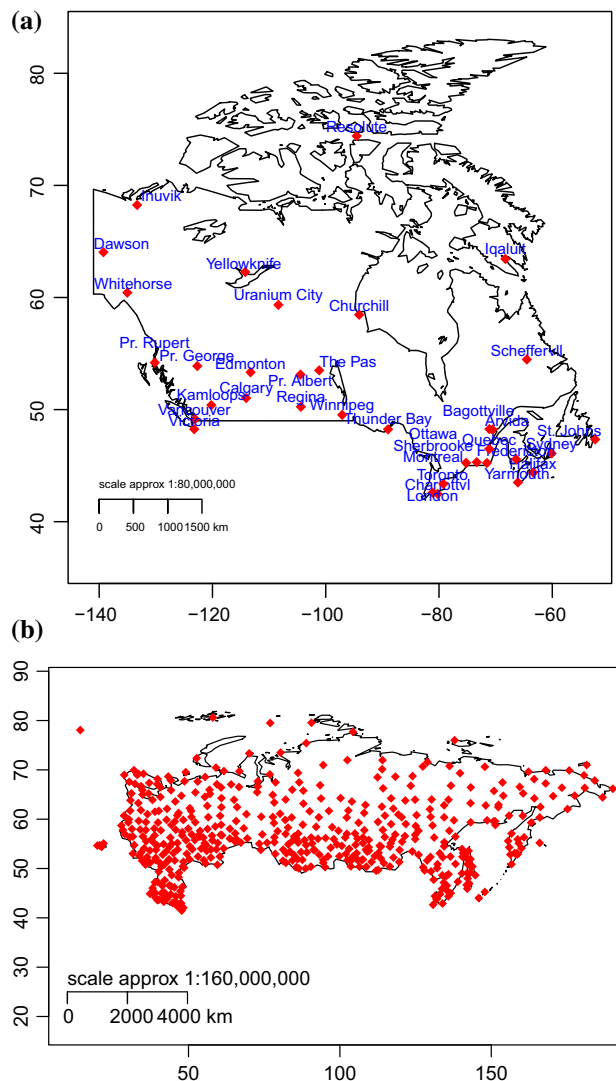


**Fig. 5** Curves of real datasets. **a** Canadian temperature data, **b** Russian temperature data, **c** radar data

years for one station. Each row is considered as a curve over interval  $[1, 365]$ . Thus, we have 490 curves that are shown in Fig. 5b. We also show the geographical locations of these data in Fig. 6b.

3. **Radar:**<sup>2</sup> This dataset is recorded by satellite Topex/Poseidon around a region of 25 kilometers on the Amazon river and contains  $N = 472$  waveforms (i.e. curves). Each curve is measured at 70 points,  $\{X_i(t_1), \dots, X_i(t_{70})\}_{i=1}^{472}$ , and corresponds to a type of ground (i.e. its natural features). To explore hydrological aims of Amazonian basin, this dataset is analyzed. The dataset forms a  $472 \times 70$  matrix which each row of this matrix is considered as a curve. These data have

<sup>2</sup> This dataset is available at <http://www.math.univ-toulouse.fr/staph/npfda/>.



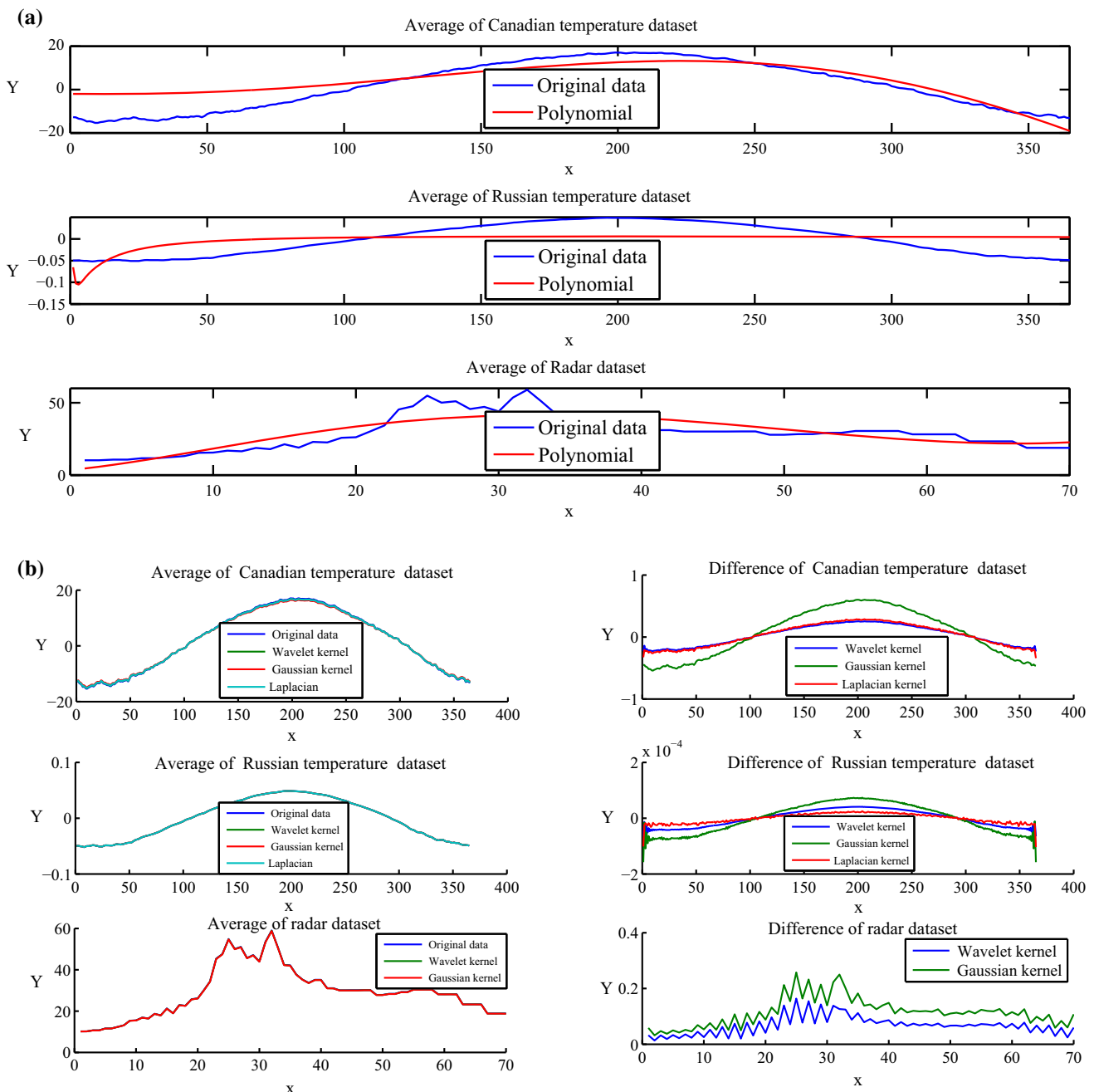
**Fig. 6** Geographical locations. **a** Location of Canadian weather stations, **b** location of Russian weather stations

clustered by Luz-López-García et al. (2015). The data are shown in Fig. 5c.

### 5.2 Cross-validation and condition number for real datasets

We determine the suitable parameters of stated kernels by CV technique. For example, to select the parameters of each kernel for Canadian temperature dataset, we calculate the average of corresponding projections (i.e. we project the data onto RKHS) for different values of  $\gamma_s$  and  $\sigma_t$  with  $s = 1, \dots, 4$ ,  $t = 1, 2, 4$  and  $\sigma_3 = (a, b)^T$ . Here  $\sigma_3$  is a vector. The optimal parameters are chosen such that the sum of square error,  $\sum_{j=1}^{35} \sum_{i=1}^{365} (f^j(x_i) - \hat{f}^{*j}(x_i))^2$ , is





**Fig. 7** a Average of projections and average of original curves via polynomial kernel for real datasets, and b average of projections and average of original curves via each kernel for real datasets. The

second column shows the differences between the average of original signals and the average of their projections for different kernels

minimized where  $f^j$  and  $\hat{f}^{*j}$  are original signal and its reconstruction, respectively [ $\hat{f}^{*j}$  is defined in Eq. (5)]. In addition, these parameters should not yield an ill-posed system. For Canadian temperature dataset, we consider the following values to find the optimal parameters,  $\gamma_s = 10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}$  and  $\sigma_t = 0.1, 0.2, \dots, 2$   $s = 1, \dots, 4$ ,  $t = 1, 2, 4$  and  $\sigma_3 = (10^{-3}, 3)^T, (10^{-3}, 4)^T, (10^{-3}, 5)^T, (10^{-4}, 3)^T, (10^{-4}, 4)^T, (10^{-4}, 5)^T, (10^{-5}, 3)^T,$

$(10^{-5}, 4)^T, (10^{-5}, 5)^T$ . Using these parameters, we calculate the kernel matrix  $K_{365 \times 365}$ , expansion coefficients  $\alpha$  and reconstruction of curves from  $\hat{f}_n^{*j} = K_x \alpha$ . We compare the obtained square errors and choose those parameters that yield minimal square errors and well-posed system in Eq. (4). This procedure can be used for Russian temperature and radar datasets. The polynomial kernel cannot be used since it yields an ill-posed system in Eq. (4) for real

**Table 4** The condition number of real datasets for polynomial kernel and  $T = (n\gamma\mathbf{I}_n + K_x)$

	Canadian temperature	Russian temperature	Radar
parameter	$\{\gamma = 1, a = 10^{-4}\}$ $b = 5$	$\{\gamma = 10^{-3}, a = 10^{-2}\}$ $b = 0.75$	$\{\gamma = 10^{-3}, a = 10^{-3}\}$ $b = 5$
$\text{cond}(K_x)$	$1.11 \times 10^{23}$	$6.99 \times 10^{19}$	$4.49 \times 10^{19}$
$\text{cond}(T)$	$6 \times 10^4$	$9.85 \times 10^4$	$8.42 \times 10^5$

datasets. In other words, if for a kernel the condition number of matrix  $K_x$  and  $T = (n\gamma\mathbf{I}_n + K_x)$  from Eq. (4) is increased, that kernel does not present a suitable reconstruction of data. In Fig. 7a, we present the average of original data and average of their reconstruction based on the polynomial kernel. For real datasets, the condition number of the polynomial kernel is presented in Table 4. As we can see in Table 4, the condition number of the polynomial kernel is very large, and it yields an ill-posed system.

The best values for the real datasets and corresponding error are shown in Table 5 where the smallest error is bolded.

Now, we represent the average of original signals and the average of their projections on RKHS (reconstruction) for each dataset in Fig. 7b. Since the average of original signals and the average of their projections are close together, we also show the differences between the average of original signals and the average of their projections in the second column. As we can see, an excellent reconstruction obtained based on wavelet kernel for Canadian temperature and radar dataset and based on Laplacian kernel for Russian temperature.

The condition number of real datasets are presented in Table 6. Gaussian kernel, wavelet kernel, and Laplacian kernel have the best condition number for Canadian temperature, Russian temperature, and radar datasets, respectively.

### 5.3 Determining the number of clusters

A major problem in clustering is the detection of the number of clusters. There are many validity indices to identify it, but they do not cover all problems. Three examples of validity indices are Dunn index, Davies–Bouldin index and Wilk’s Lambda index. We explain them in the following.

The cluster diameters and dissimilarity between of clusters are combined to obtain the most suitable number of clusters by Dunn index (Dunn 1974). The high value of Dunn index indicates that clusters are compact and well separated.

Davies–Bouldin index (Davies and Bouldin 1979) can be used to measure the dispersion of a cluster and

**Table 5** Kernel parameters for real datasets and corresponding error

Data name	Laplacian $\sigma_1, \gamma_1$	Gaussian $\sigma_2, \gamma_2$	Wavelet $\sigma_3, \gamma_4$
Canadian temperature	$1, 10^{-4}$	$0.1, 10^{-4}$	$1.4, 10^{-4}$
Error	0.0471	0.2081	<b>0.0368</b>
Russian temperature	$2, 10^{-5}$	$1.4, 10^{-5}$	$1.7, 10^{-5}$
Error	<b>6.84</b> $\times 10^{-10}$	$35.8 \times 10^{-10}$	$10.3 \times 10^{-10}$
Radar	$1, 10^{-4}$	$1, 10^{-4}$	$1.5, 10^{-4}$
Error	0.0509	0.0681	<b>0.0270</b>

**Table 6** Condition number of real datasets and  $T = (n\gamma\mathbf{I}_n + K_x)$

Datasets		Laplacian	Gaussian	Wavelet
Canadian temperature	$\text{cond}(K_x)$	4.68	<b>1</b>	7.16
	$\text{cond}(T)$	4.41	<b>1</b>	6.66
Russian temperature	$\text{cond}(K_x)$	64.59	63.05	<b>49.64</b>
	$\text{cond}(T)$	62.78	57.71	<b>48.03</b>
Radar	$\text{cond}(K_x)$	<b>4.6727</b>	5.8823	12.8054
	$\text{cond}(T)$	<b>4.6179</b>	5.7714	12.5205

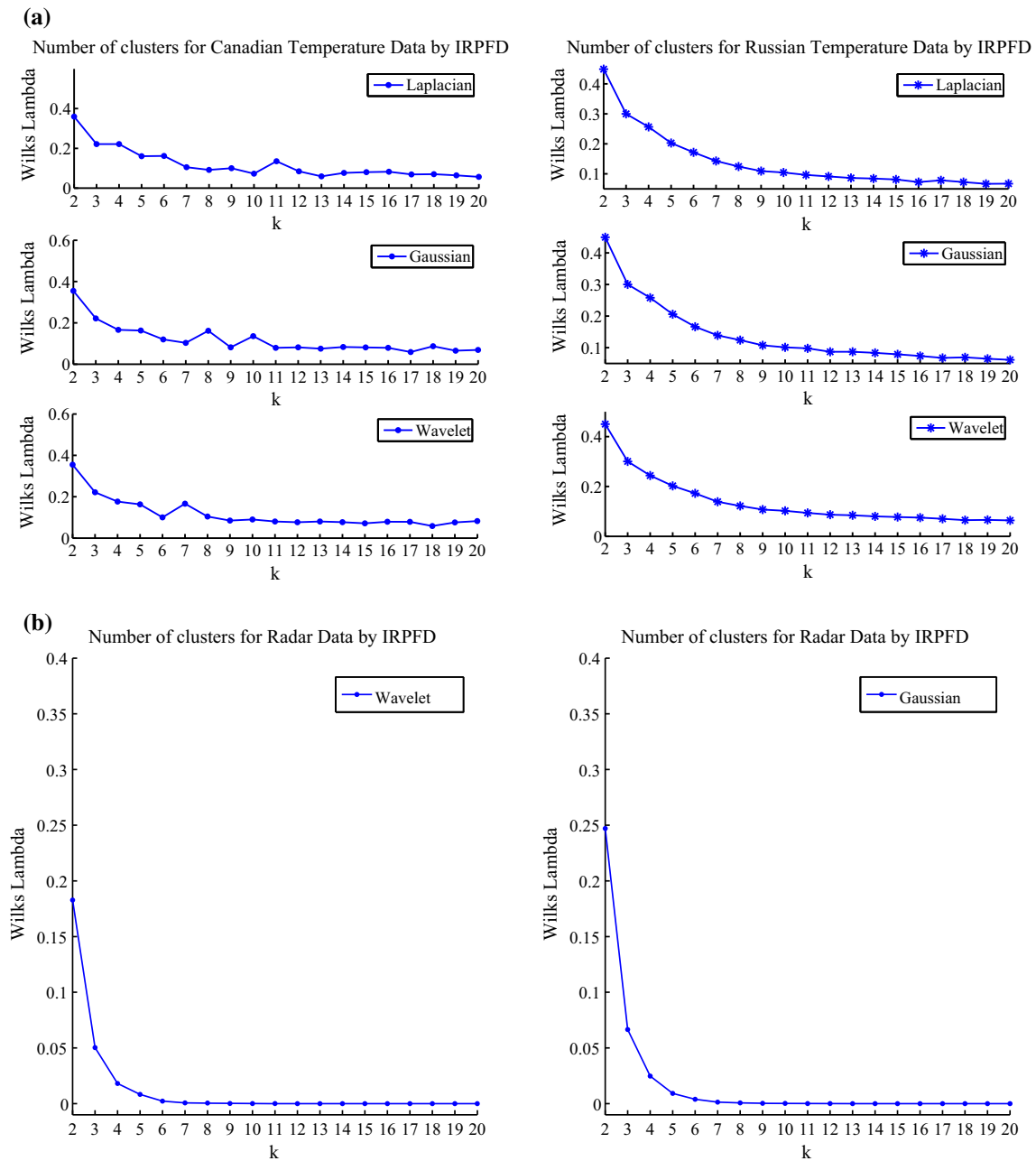
dissimilarity between clusters. The small value of Davies–Bouldin index gives the suitable number of clusters.

These indices are successful only in special cases, and they are not adequate for the general situation. Dunn index requires expensive computations and is sensible to the noise of data. It is not used for noisy data, moderate and big data. Also, Davies–Bouldin index does not give good results for overlapping clusters (Saitta et al. 2007).

For real data, the number of clusters is unknown. Canadian temperature dataset is clustered by Jacques and Preda (2014b) with 4 clusters. Luz-López-García et al. (2015) using Wilk’s Lambda (presented in the following) determine  $k = 3$  for radar dataset. However, we used the mentioned indices (Dunn index and Davies–Bouldin index) for the real data and did not get appropriate results for the number of clusters.<sup>3</sup>

Wilk’s Lambda (Kuo and Lin 2010) is usually used to determine the number of clusters. It is defined as follows:

<sup>3</sup> The results are available upon request from the authors.



**Fig. 8** **a** The values of Wilk’s Lambda for Canadian temperature dataset (the *first column*) and Russian temperature dataset (the *second column*) for different kernels, and **b** the values of Wilk’s Lambda for radar dataset

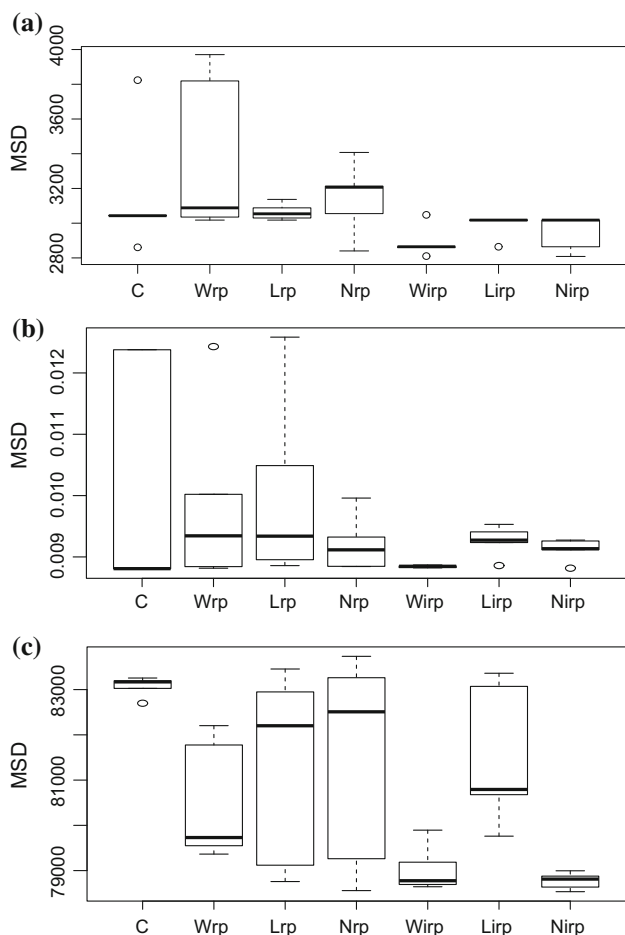
$$\text{Wilk's Lambda} = \frac{SS_{\text{within}}^k}{SS_{\text{total}}}; \quad k = 2, \dots, k_{\text{max}},$$

where  $SS_{\text{within}}^k$  is the within-cluster variance and  $SS_{\text{total}}$  is the total variance. Using IRPFD K-means, we determine it for Canadian temperature and Russian temperature datasets by three kernels and radar dataset by two kernels (Gaussian and wavelet). For  $k = 2, 3, \dots, 20$ , we present Wilk’s lambda in Fig. 8a, b for three datasets. We choose  $k = 4$ ,  $k = 8$  and  $k = 3$  for Canadian temperature dataset, Russian

temperature dataset, and radar dataset, respectively. Wilk’s lambda does not considerably decrease after the selected points based on the most kernels.

### 5.4 The setting of IRPFD for Canadian temperature dataset

Now, we explain working of IRPFD for Canadian temperature dataset by previously obtained parameters in Sects. 5.2 and 5.3. The optimal parameters are  $\sigma_1 = 1$ ,



**Fig. 9** MSD of real datasets. **a** Canadian temperature dataset, **b** Russian temperature, **c** radar dataset

$\gamma_1 = 10^{-4}$ ,  $\sigma_2 = 0.1$ ,  $\gamma_2 = 10^{-4}$ ,  $\sigma_3 = 1.4$  and  $\gamma_3 = 10^{-4}$ . The number of clusters is  $k = 4$ . Here, we have  $N = 35$ ,  $n = 365$ ,  $d_z = (10, 40, 60, 80)^T$  and  $X_n = \{1, 2, \dots, 365\}$ . Matrix  $K$  is computed for three kernels using the mentioned values. Here, we consider Gaussian kernel which is as

$$K(x_i, x_j) = e^{-\frac{1}{0.1^2} \|x_i - x_j\|_2^2}$$

where  $\|\cdot\|_2$  is Euclidean norm and  $i, j = 1, \dots, 365$ . We obtain its Cholesky decomposition as  $K = UU^T$  where matrix  $U$  is  $365 \times 365$ . For  $j = 1, \dots, 35$ , we solve the linear system Eq. (4) to obtain  $\alpha^j$  for Gaussian kernel. Also,  $\alpha^j$  is a vector of size  $365 \times 1$ . We generate a random matrix  $R_{10 \times 365}$  which elements are drawn from the standard normal distribution and set  $P_{10 \times 365} = R_{10 \times 365} U_{365 \times 365}$ . For  $j = 1, \dots, 35$ , we project expansion coefficients by RP as  $\beta_{10 \times 1}^j = P_{10 \times 365} \alpha_{365 \times 1}^j$  and put them in the rows of a matrix  $\mathcal{A}_{35 \times 10}^{RP_1}$ . Now, we initialize  $G$  by generating 35 random numbers from  $unif(\{1, 2, 3, 4\})$  where  $unif$  is discrete uniform distribution. According to  $G$ , we calculate the

mean of each cluster and put them in the rows of a matrix  $C_{4 \times 10}^{RP_1}$  as cluster centers. That means

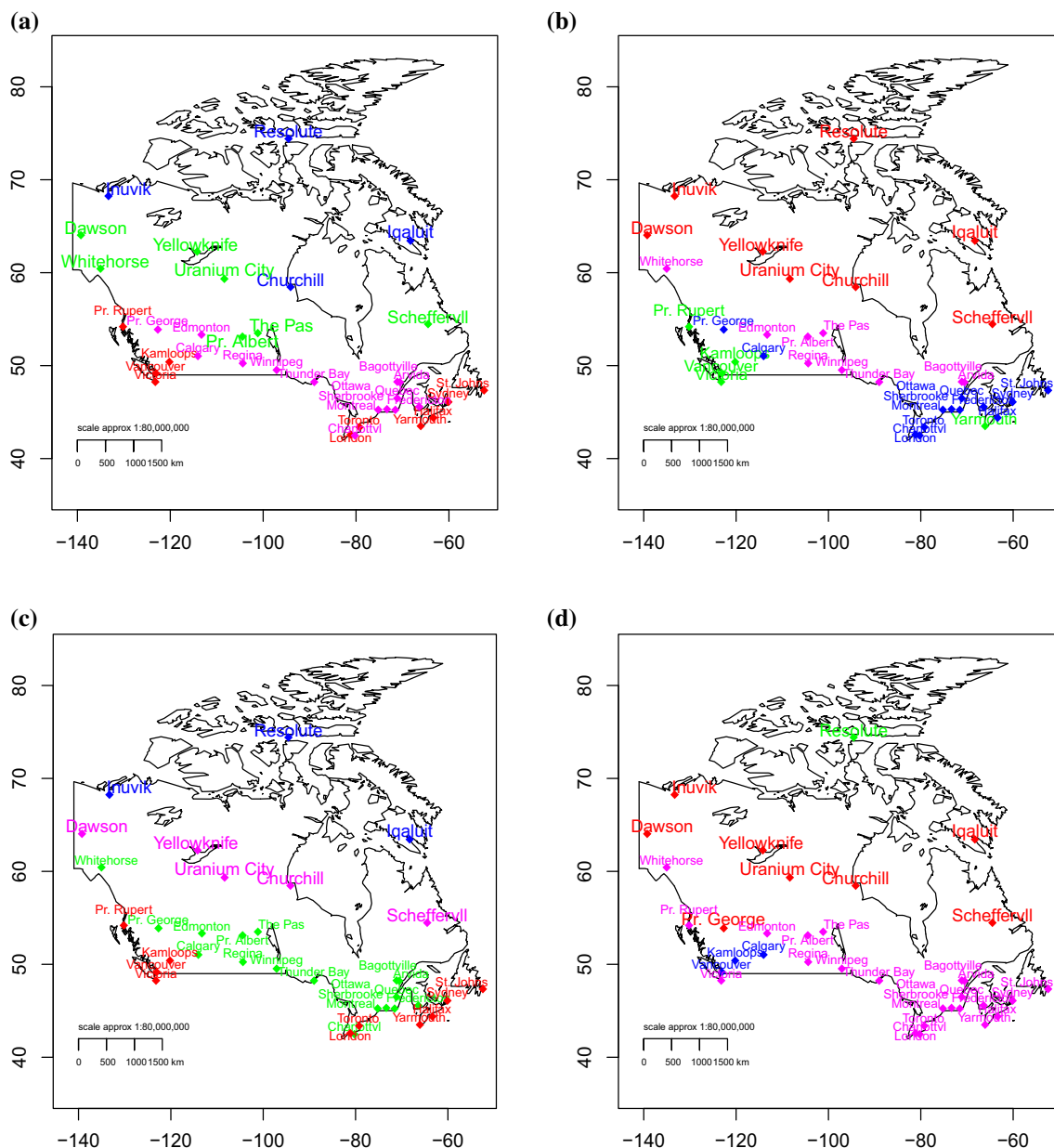
$$C_{4 \times 10}^{RP_1} = \begin{bmatrix} \frac{1}{|\{i|g_i = 1\}|} \sum_{\{i|g_i = 1\}} A_i \\ \frac{1}{|\{i|g_i = 2\}|} \sum_{\{i|g_i = 2\}} A_i \\ \frac{1}{|\{i|g_i = 3\}|} \sum_{\{i|g_i = 3\}} A_i \\ \frac{1}{|\{i|g_i = 4\}|} \sum_{\{i|g_i = 4\}} A_i \end{bmatrix},$$

where  $|\cdot|$  is the cardinality of a set,  $g_i$  belongs to  $G$  and  $A_i$  is  $i$ -th row of  $\mathcal{A}_{35 \times 10}^{RP_1}$  for  $i = 1, \dots, 35$ . We cluster  $\mathcal{A}_{35 \times 10}^{RP_1}$  by K-means which is initialized with  $C_{4 \times 10}^{RP_1}$ . K-means gives a new  $G$ . We generate a new random matrix  $P_{40 \times 365}$  as previous and obtain  $\mathcal{A}_{35 \times 40}^{RP_2}$ . Using the obtained  $G$  and  $\mathcal{A}_{35 \times 40}^{RP_2}$ , we calculate  $C_{4 \times 40}^{RP_2}$ . Applying K-means on  $\mathcal{A}_{35 \times 40}^{RP_2}$  using  $C_{4 \times 40}^{RP_2}$  gives a new  $G$ . We perform these process for dimension 60 and 80. Using the obtained  $G$  in dimension 80, we calculate the cluster centers in the original space.  $d_z = (10, 40, 60, 80)^T$  is selected from candidate values which are minimized MSD.

### 5.5 Results

In this section, we present the results of applying classical K-means, RPF and IRPF on real datasets. Note that we use hard clustering techniques. However, there are several works in hydrometeorology which used fuzzy clustering and find statistically and climatologically homogeneous regions, (Srinivas 2008; Satyanarayana and Srinivas 2011; Bharath and Srinivas 2015a). The clustering results of real datasets (similar to those of simulated datasets) over 5 iterations are shown in Table 8 where SD is a standard deviation. Note that for real datasets, Rand index cannot be calculated since the true clusters of the real datasets are unknown. The proposed method with wavelet kernel for the first two datasets and with the Gaussian kernel for the third dataset has the best MSD among other methods. Let us note that Jacques and Preda (2014b) also clustered Canadian temperature and precipitation data. For their obtained clusters, we calculate the MSD of the temperature data. It is equal to  $3.56 \times 10^3$  which is greater than MSD of the proposed method ( $2.89 \times 10^3$ ).

The boxplots for MSD are shown in Fig. 9. According to this figure, IRPF based on wavelet kernel for Canadian temperature and Russian temperature datasets and based on



**Fig. 10** Geographical locations of Canadian temperature dataset with **a** classical K-means, **b** RPF D, **c** IRPF D based on wavelet kernel and **d** the results of Jacques and Preda (2014b)

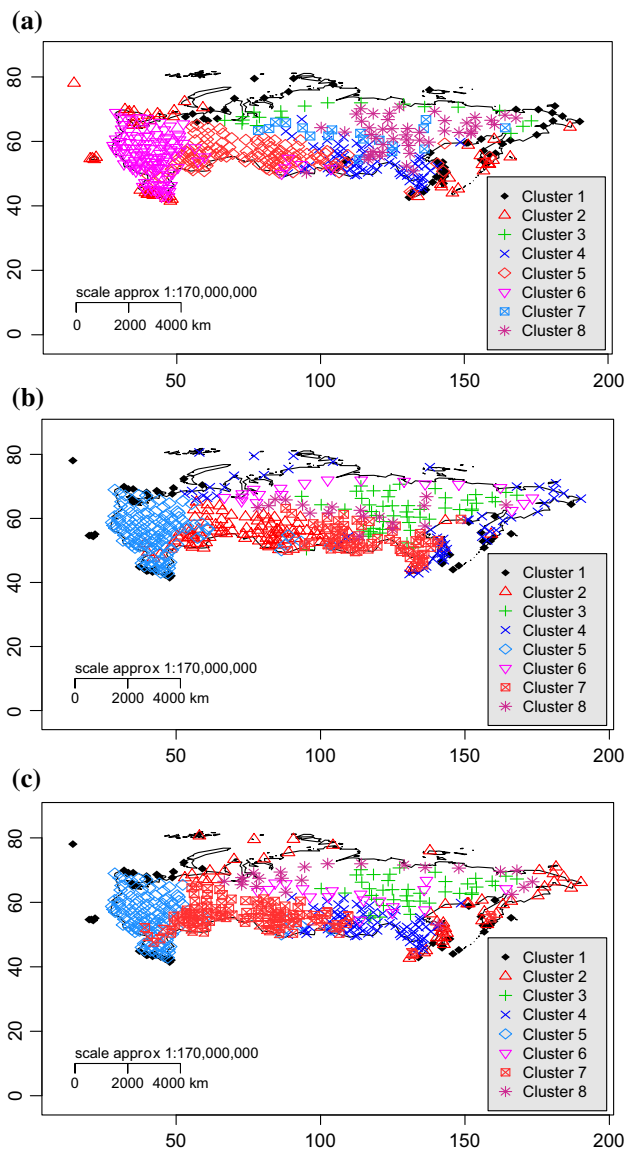
the Gaussian kernel for radar dataset has the best performance among others.

The results of geographical locations of Canadian temperature dataset for classical K-means, RPF D and IRPF D based on wavelet kernel are shown in Fig. 10. We also present the geographical locations of the results which are found by Jacques and Preda (2014b) in Fig. 10d. Detected clusters by IRPF D and Jacques and Preda (2014b) are different for northern and southern stations, obviously.

The geographical locations of Russian temperature dataset for classical K-means, RPF D and IRPF D based on wavelet kernel are shown in Fig. 11. According to

Figs. 10c and 11c, IRPF D has founded the homogeneous clusters for Canadian temperature and Russian temperature datasets which are consistent with the reality of climate of Canada and Russia. Note that we do not use the geographical positions of the stations to obtain these clusters for Canadian temperature and Russian temperature datasets. The proposed method is only for univariate functional random variable. Horenko (2010) also clustered the historical temperature data in Europe only using the temperature. If we consider temperature and precipitation variables, then we have a bivariate functional data. Let us note that Satyanarayana and Srinivas (2011) and Bharath





**Fig. 11** Geographical locations of Russian temperature dataset with **a** classical K-means, **b** RPFDF and **c** IRPFDF based on wavelet kernel

et al. (2016) proposed an approach for regionalization of precipitation and temperature, respectively using location indicators (e.g. latitude, longitude, altitude and distance to sea) and large-scale atmospheric variables and they obtained suitable results.

The curves of each cluster which obtained with IRPFDF based on wavelet kernel for Canadian temperature dataset and based on the Gaussian kernel for radar dataset are presented in Fig. 12a, b. For Canadian temperature dataset with IRPFDF, the cluster 1 contains 17 signals, the cluster 2 has 10 signals, the cluster 3 has 3 signals and the fourth cluster has 5 signals while Jacques and Preda (2014b) have found 23, 3, 1 and 8 signals for the clusters 1, 2, 3 and 4, respectively. For radar data using IRPFDF, the clusters 1, 2 and 3 have 88, 50

and 334 signals while those obtained by Luz-López-García et al. (2015) are 94, 47 and 329, respectively.

For radar dataset, we present the details of running time for classical K-means, RPFDF and IRPFDF based on the Gaussian kernel in Table 7. According to Algorithms 1 and 2, the running time of RPFDF and IRPFDF is the sum of the running time of: calculate kernel matrix (denoted by  $t_{ker}$ ), Cholesky decomposition (denoted by  $t_{Ch}$ ), compute coefficients  $\alpha$  (denoted by  $t_{alp}$ ) and K-means (denoted by  $t_K$ ). According to this table, the K-means running time ( $t_K$ ) is fairly small compared to the total running time.

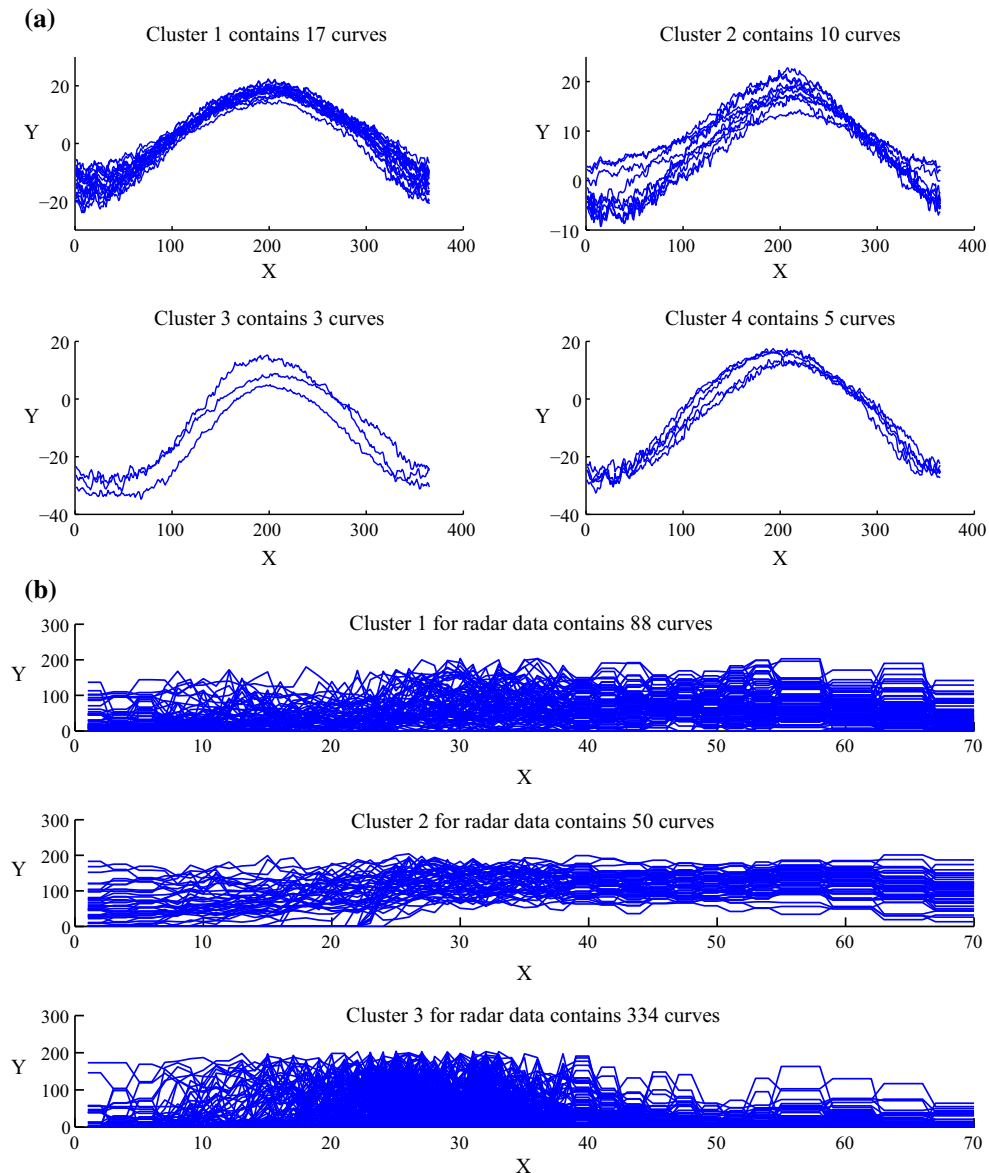
### 6 Conclusions

Although functional data clustering can be performed by clustering raw data (without considering the functional nature), there are examples of datasets that we must consider the functional nature to obtain the appropriate results. In other words, functional data clustering presents a better result than classical K-means (for example see MSD of radar dataset in Table 8). In this paper, we have introduced an iterative random projection K-means for functional data (IRPFDF) which is a method to cluster functional data based on random projection (RP) and reproducing kernel Hilbert space (RKHS). We increase the dimension several times to avoid local minimum.

We implement three methods (classical K-means, RPFDF, IRPFDF) on some simulated and climatological datasets. For simulation study, Rand index (RI) and mean square distance (MSD) of IRPFDF is the best in all datasets, and its average number of iterations (Iter) is the best for the most datasets. In general, clustering via functional nature is better than classical K-means (in some dataset classical K-means is better than RPFDF based on RI and MSD). The running time of classical K-means is less than others because it uses K-means only one time on raw data. This is while the running time of IRPFDF K-means and RPFDF K-means are the same approximately in the most datasets (i.e. this is not for all datasets). Thus, IRPFDF takes a little longer time than classical K-means and RPFDF to find the best solution, and it has a smaller Iter than others.

We have used the wavelet kernel that gives a better reconstruction of signals based on Tikhonov regularization theory for the most datasets. This kernel also improves the Iter, RI and MSD of RPFDF and IRPFDF in the most cases.

For real dataset, we determine the number of cluster using Wilk’s Lambda which is based on the within-cluster variance. Since the true clusters of real dataset are unknown, Rand index is not calculated. For Canadian temperature and Russian temperature datasets, IRPFDF based on wavelet kernel finds a solution with smallest MSD and for radar dataset based on the Gaussian kernel.



**Fig. 12** **a** The solution of IRPFD based on wavelet kernel for Canadian temperature dataset, and **b** the solution of IRPFD based on Gaussian kernel for radar dataset

**Table 7** The details of the running time for radar dataset

Methods	Total running time	t.ker	t.Ch	t.alp	t.K
Classical K-means	0.0589				
RPFD	0.0845	0.0333	$1.8027 \times 10^{-4}$	0.0186	0.0325
IRPFD	0.0842	0.0333	$1.8027 \times 10^{-4}$	0.0186	0.0321

To cluster functional data, we project data onto RKHS, and we have to determine the suitable parameters of Mercer kernel and the regularization parameter (in this paper, we have 4 Mercer kernels). It is not clear which kernel is appropriate for data before clustering. Determining a true number of clusters and an optimal dimension for RP is a hard task. Future research can be

done to present another guideline for kernel parameterization, detecting the true number of clusters and an optimal dimension for RP. The proposed method can be extended to multivariate functional data and spatial data in future works. The development of fuzzy framework of the proposed method could be considered in a future research.

**Table 8** The clustering results of real data: Time, Iter, MSD. Also, SD is a standard deviation

Datasets		Classical	RPF			IRPF		
			Laplacian	Gaussian	Wavelet	Laplacian	Gaussian	Wavelet
Canadian temperature	Iter	3.40	1.80	2.80	3.80	<b>0.40</b>	0.80	0.60
	SD	2.30	1.30	2.17	1.92	0.55	1.30	1.34
$h = 80$	Time (s)	<b>0.00</b>	0.16	0.36	0.19	0.16	0.37	0.19
	SD	0.01	0.00	0.00	0.00	0.00	0.00	0.00
$d_z = (10, 40, 60, 80)^T$	MSD ( $\times 10^3$ )	3.16	3.06	3.38	3.18	2.94	2.99	<b>2.89</b>
	SD	377.8	211.1	468.2	48.3	101.3	91.4	68.6
Russian temperature	Iter	14.8	17	18.2	24.2	<b>3.8</b>	4.4	6.2
	SD	1.9	5.9	8.7	13.6	4.1	3.4	7.8
$h = 110$	Time (s)	<b>0.05</b>	0.45	0.42	0.79	0.46	0.43	0.80
	SD ( $\times 10^{-4}$ )	219	52	84	130	50	69	77
$d_z = (10, 50, 100, 110)^T$	MSD ( $\times 10^{-4}$ )	102	100	92	99	93	91	<b>88</b>
	SD ( $\times 10^{-4}$ )	20	16	5	15	3	2	0
Radar	Iter	12.60	10.20	16.00	13.00	6.20	<b>4.20</b>	5.60
	SD	9.91	3.96	11.31	7.39	3.49	4.44	0.89
$h = 45$	Time (s)	<b>0.01</b>	0.02	0.03	0.06	0.03	0.04	0.05
	SD	0.02	0.00	0.00	0.01	0.01	0.00	0.01
$d_z = (10, 15, 30, 45)^T$	MSD ( $\times 10^4$ )	8.31	8.13	8.15	8.05	8.15	<b>7.88</b>	7.90
	SD ( $\times 10^3$ )	0.22	2.20	2.39	1.35	1.59	0.19	0.52

**Acknowledgements** The authors gratefully acknowledge the useful comments and suggestions of associate editor and two anonymous reviewers which help us to improve the paper.

**References**

Abraham C, Cornillon PA, Matzner-Løber E, Molinari N (2003) Unsupervised curve clustering using B-splines. *Scand J Stat* 30(3):581–595

Abramowicz K, Arnqvist P, Secchi P, Luna S, Vantini S, Vitelli V (2017) Clustering misaligned dependent curves applied to varved lake sediment for climate reconstruction. *Stoch Environ Res Risk Assess* 31(1):71–85. doi:10.1007/s00477-016-1287-6

Aguilera-Morillo MC, Durban M, Aguilera AM (2017) Prediction of functional data with spatial dependence: a penalized approach. *Stoch Environ Res Risk Assess* 31(1):7–22. doi:10.1007/s00477-016-1216-8

Akila Y (1999) A hierarchical approach for the regionalization of precipitation annual maxima in Canada. *J Geophys Res Atmos* 104(24):31,645–31,655

Antoniadis A, Brossat X, Cugliari J, Poggi JM (2013) Clustering functional data using wavelets. *Int J Wavelets Multiresolution Inf Process* 11(1):1350003

Anyadike RNC (1987) A multivariate classification and regionalization of West African climates. *J Climatol* 7(2):157–164

Aroszajn N (1950) Theory of reproducing kernels. *Trans Am Math Soc* 68(3):337–404

Asong ZE, Khaliq MN, Wheeler HS (2015) Regionalization of precipitation characteristics in the Canadian prairie provinces using large-scale atmospheric covariates and geophysical attributes. *Stoch Environ Res Risk Assess* 29(3):875–892. doi:10.1007/s00477-014-0918-z

Brring L (1988) Regionalization of daily rainfall in Kenya by means of common factor analysis. *J Climatol* 8(4):371–389. doi:10.1002/joc.3370080405

Bagirov AM, Mardaneh K (2006) Modified global k-means algorithm for clustering in gene expression data sets. In: Boden M, Bailey TL (eds) 2006 Workshop on intelligent systems for bioinformatics (WISB 2006). ACS, Hobart, CRPIT, pp 23–28

Balzanella A, Romano E, Verde R (2017) Modified half-region depth for spatially dependent functional data. *Stoch Environ Res Risk Assess* 31(1):87–103. doi:10.1007/s00477-016-1291-x

Bernardi MS, Sangalli LM, Mazza G, Ramsay JO (2017) A penalized regression model for spatial functional data with application to the analysis of the production of waste in Venice province. *Stoch Environ Res Risk Assess* 31(1):23–38. doi:10.1007/s00477-016-1237-3

Bharath R, Srinivas VV (2015a) Delineation of homogeneous hydrometeorological regions using wavelet-based global fuzzy cluster analysis. *Int J Climatol* 35(15):4707–4727

Bharath R, Srinivas VV (2015b) Regionalization of extreme rainfall in India. *Int J Climatol* 35(6):1142–1156. doi:10.1002/joc.4044

Bharath R, Srinivas VV, Basu B (2016) Delineation of homogeneous temperature regions: a two-stage clustering approach. *Int J Climatol* 36(1):165–187. doi:10.1002/joc.4335

Bohorquez M, Giraldo R, Mateu J (2017) Multivariate functional random fields: prediction and optimal sampling. *Stoch Environ Res Risk Assess* 31(1):53–70. doi:10.1007/s00477-016-1266-y

Boullé M (2012) Functional data clustering via piecewise constant nonparametric density estimation. *Pattern Recogn* 45(12):4389–4401

Bouveyron C, Jacques J (2011) Model-based clustering of time series in group-specific functional subspaces. *Adv Data Anal Classif* 5(4):281–300

Bulygina ON, Razuvaev VN (2012) Daily temperature and precipitation data for 518 Russian meteorological stations. Carbon Dioxide Information Analysis Center, Oak Ridge National Laboratory, US Department of Energy, Oak Ridge, Tennessee

- Cardoso A, Wichert A (2012) Iterative random projections for high-dimensional data clustering. *Pattern Recogn Lett* 33(13):1749–1755
- Dasgupta S, Gupta A (2003) An elementary proof of a theorem of Johnson and Lindenstrauss. *Random Struct Algorithm* 22:60–65
- Davies DL, Bouldin DW (1979) A cluster separation measure. *IEEE Trans Pattern Anal Mach Intell PAMI* 1(2):224–227
- Delaigle A, Hall P (2010) Defining probability density for a distribution of random functions. *Ann Stat* 38:1171–1193
- Dunn JC (1974) Well-separated clusters and optimal fuzzy partitions. *J Cybern* 4(1):95–104. doi:[10.1080/01969727408546059](https://doi.org/10.1080/01969727408546059)
- El-Jabi N, Ashkar F, Hebabi S (1998) Regionalization of floods in New Brunswick (Canada). *Stoch Hydrol Hydraul* 12(1):65–82. doi:[10.1007/s004770050010](https://doi.org/10.1007/s004770050010)
- Everitt BS, Landau S, Leese M, Stahl D (2011) *Cluster analysis*. Wiley series in probability and statistics. Kings College London, London
- Ferraty F, Vieu P (2006) *Nonparametric functional data analysis*. Springer series in statistics. Springer, New York
- Ferreira L, Hitchcock DB (2009) A comparison of hierarchical methods for clustering functional data. *Commun Stat Simul Comput* 38:1925–1949
- Finazzi F, Haggarty R, Miller C, Scott M, Fassò A (2015) A comparison of clustering approaches for the study of the temporal coherence of multiple time series. *Stoch Environ Res Risk Assess* 29(2):463–475
- Giacofci M, Lambert-Lacroix S, Marot G, Picard F (2013) Wavelet-based clustering for mixed-effects functional models in high dimension. *Biometrics* 69:31–40
- Giuseppe ED, Lasinio GJ, Esposito S, Pasqui M (2013) Functional clustering for Italian climate zones identification. *Theoret Appl Climatol* 114(1):39–54
- González-Hernández J (2010) Representing functional data in reproducing kernel Hilbert spaces with applications to clustering, classification and time series problems. Ph.D. thesis, Department of Statistics, Unisversidad Carlos III, Getafe, Madrid
- Guenni L, Bárdossy A (2002) A two steps disaggregation method for highly seasonal monthly rainfall. *Stoch Environ Res Risk Assess* 16(3):188–206
- Haggarty R, Miller C, Scott E, Wyllie F, Smith M (2012) Functional clustering of water quality data in scotland. *Environmetrics* 23(8):685–695
- Horenko I (2010) On clustering of non-stationary meteorological time series. *Dyn Atmos Oceans* 49(23):164–187. doi:[10.1016/j.dynatmoce.2009.04.003](https://doi.org/10.1016/j.dynatmoce.2009.04.003)
- Hubert L, Arabie P (1985) Comparing partitions. *J Classif* 2(1):193–218
- Ieva F, Paganoni AM, Pigoli D, Vitelli V (2013) Multivariate functional clustering for the morphological analysis of electrocardiograph curves. *J R Stat Soc Ser C* 62(3):401–418
- Jacques J, Preda C (2013) Funclust: a curves clustering method using functional random variable density approximation. *Neurocomputing* 112:164–171
- Jacques J, Preda C (2014a) Functional data clustering: a survey. *Adv Data Anal Classif* 8:231–255
- Jacques J, Preda C (2014b) Model-based clustering for multivariate functional data. *Comput Stat Data Anal* 71:92–106
- James G, Sugar C (2003) Clustering for sparsely sampled functional data. *J Am Stat Assoc* 98(462):397–408
- Johnson W, Lindenstrauss J (1984) Extensions of Lipschitz mappings into a Hilbert space. *Contemp Math* 26:189–206
- Jolliffe IT (2002) *Principal component analysis*, 2nd edn. Springer, New York
- Kuo RJ, Lin LM (2010) Application of a hybrid of genetic algorithm and particle swarm optimization algorithm for order clustering. *Decis Support Syst* 49:451–462
- Luz-López-García M, García-Ródenas R, González-Gómez A (2015) K-means algorithms for functional data. *Neurocomputing* 151(1):231–245
- Mateu J, Romano E (2017) Advances in spatial functional statistics. *Stoch Environ Res Risk Assess* 31(1):1–6. doi:[10.1007/s00477-016-1346-z](https://doi.org/10.1007/s00477-016-1346-z)
- Mitchell VL (1976) The regionalization of climate in the western United States. *J Appl Meteorol* 15(9):920–927. doi:[10.1175/1520-0450\(1976\)015<0920:TROCIT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1976)015<0920:TROCIT>2.0.CO;2)
- Nam W, Shin H, Jung Y, Joo K, Heo JH (2015) Delineation of the climatic rainfall regions of South Korea based on a multivariate analysis and regional rainfall frequency analyses. *Int J Climatol* 35(5):777–793. doi:[10.1002/joc.4182](https://doi.org/10.1002/joc.4182)
- Rahman A, Charron C, Ouarda TBMJ, Chebana F (2017) Development of regional flood frequency analysis techniques using generalized additive models for Australia. *Stoch Environ Res Risk Assess*. doi:[10.1007/s00477-017-1384-1](https://doi.org/10.1007/s00477-017-1384-1)
- Ramsay J, Silverman B (2005) *Functional data analysis*. Springer series in statistics, 2nd edn. Springer, New York
- Ramsay JO (1982) When the data are functions. *Psychometrika* 47(4):379–396. doi:[10.1007/BF02293704](https://doi.org/10.1007/BF02293704)
- Rand WM (1971) Objective criteria for the evaluation of clustering methods. *J Am Stat Assoc* 66(336):846–850
- Rao AR, Srinivas V (2006a) Regionalization of watersheds by fuzzy cluster analysis. *J Hydrol* 318(14):57–79. doi:[10.1016/j.jhydrol.2005.06.004](https://doi.org/10.1016/j.jhydrol.2005.06.004)
- Rao AR, Srinivas V (2006b) Regionalization of watersheds by hybrid-cluster analysis. *J Hydrol* 318(14):37–56. doi:[10.1016/j.jhydrol.2005.06.003](https://doi.org/10.1016/j.jhydrol.2005.06.003)
- Ray S, Mallick B (2006) Functional clustering by Bayesian wavelet methods. *J R Stat Soc B* 68(2):305–332
- Rossi F, Conan-Guez B, Golli AE (2004) Clustering functional data with the SOM algorithm. *Proc ESANN* 2004:305–312
- Ruiz-Medina MD, Espejo RM (2012) Spatial autoregressive functional plug-in prediction of ocean surface temperature. *Stoch Environ Res Risk Assess* 26(3):335–344. doi:[10.1007/s00477-012-0559-z](https://doi.org/10.1007/s00477-012-0559-z)
- Ruiz-Medina MD, Espejo RM, Ugarte MD, Militino AF (2014) Functional time series analysis of spatio-temporal epidemiological data. *Stoch Environ Res Risk Assess* 28(4):943–954. doi:[10.1007/s00477-013-0794-y](https://doi.org/10.1007/s00477-013-0794-y)
- Saitta S, Raphael B, Smith IFC (2007) A bounded index for cluster validity. In: *Proceedings of the international conference on machine learning and data mining in pattern recognition*. Springer, Berlin, pp 174–187
- Satyanarayana P, Srinivas V (2011) Regionalization of precipitation in data sparse areas using large scale atmospheric variables a fuzzy clustering approach. *J Hydrol* 405(34):462–473. doi:[10.1016/j.jhydrol.2011.05.044](https://doi.org/10.1016/j.jhydrol.2011.05.044)
- Selim SZ, Alsultan K (1991) A simulated annealing algorithm for the clustering problem. *Pattern Recogn* 24:1003–1008
- Srinivas V, Tripathi S, Rao AR, Govindaraju RS (2008) Regional flood frequency analysis by combining self-organizing feature map and fuzzy clustering. *J Hydrol* 348(12):148–166. doi:[10.1016/j.jhydrol.2007.09.046](https://doi.org/10.1016/j.jhydrol.2007.09.046)
- Tikhonov A, Arsenin VY (1997) *Solutions of ill-posed problems*. Wiley, New York
- Tokushige S, Yadohisa H, Inada K (2007) Crisp and fuzzy k-means clustering algorithms for multivariate functional data. *Comput Stat* 22(1):1–16
- Trefethen LN, Bau D (1997) *Numerical linear algebra*. SIAM, Washington
- Yamamoto M (2012) Clustering of functional data in a low-dimensional subspace. *Adv Data Anal Classif* 6:219–247