Int J Data Sci Anal (2017) 4:301–312
DOI 10.1007/s41060-017-0069-7

REGULAR PAPER

A spectral clustering approach for multivariate geostatistical data

Francky Fouedjio[1]

Received: 5 February 2017 / Accepted: 29 August 2017 / Published online: 7 September 2017

Abstract Spectral clustering has recently become one of the most popular modern clustering methods for conventional data. However, applied to geostatistical data, the general spectral clustering method produces clusters that are spatially non-contiguous which is certainly undesirable for many geoscience applications. In this paper, a spectral clustering approach is proposed, allowing to discover spatially contiguous and meaningful clusters in multivariate geostatistical data, in which spatial dependence plays an important role. The proposed spectral clustering approach relies on a similarity measure built from a nonparametric kernel estimator of the multivariate spatial dependence structure of the data, emphasizing the spatial correlation among data locations. It integrates existing methods to find the relevant number of clusters and to assess the contribution of variables in the formation of the clusters. The results from both synthetic and real-world datasets demonstrate that the proposed spectral clustering approach can effectively provide spatially contiguous and meaningful clusters.

Keywords Geostatistics · Spectral clustering · Spatial dependency · Spatial contiguity

The present paper is an extension version of the ADMA2016 accepted paper "Discovering Spatially Contiguous Clusters in Multivariate Geostatistical Data Through Spectral Clustering" [18]. It includes an exhaustive experimental study and a new real case study.

✉ Francky Fouedjio
francky.fouedjiokameni@csiro.au;
fouedjiofrancky@yahoo.fr; francky.fouedjio@gmail.com

[1] CSIRO Mineral Resources, 26 Dick Perry Ave, Kensington, WA 6151, Australia

1 Introduction

The study and understanding of spatial phenomena in geosciences often depends on the analysis of multivariate geostatistical data, namely multivariate spatially indexed data where the indexing is continuous across space. However, such type of data poses substantial analysis challenges. One of them is the clustering of data locations into spatially contiguous clusters so that data locations in the same cluster are similar to each other and different from those in other clusters. Some applications in geosciences are [42]: (i) defining climate zones, (ii) defining coastal zone environments, (iii) defining ore typologies, (iv) identifying areas of similar land use, and (v) identifying hazardous waste sites.

In recent years, spectral clustering has become one of the most popular modern clustering methods for classical data [15,31,35,36,41]. This clustering method relies on the eigen-decomposition of a feature similarity matrix to partition observations into disjoint clusters, while considering observations in the same cluster having high similarity and observations in different clusters having low similarity. Advantages of using spectral clustering include its flexibility in terms of incorporating diverse types of similarity measures, the superiority of its clustering solution compared to traditional clustering methods such as k-means, and its well-established theoretical properties [9,25,32,33,49].

However, the application of the general spectral clustering method to geostatistical data has a tendency to produce spatially scattered clusters, which is certainly undesirable for many geoscience applications. By assuming the independence of observations, this clustering method is not able to produce spatially contiguous clusters. This fundamental assumption, however, is no longer valid in the realm of geostatistical data. Geostatistical data differs from conventional data because they often exhibit properties of spatial

dependency over the study spatial domain. This means that observations located close to one another in the geographical domain tend to have similar characteristics. In addition, the mean, the variance, and the spatial dependence structure can be different from one spatial subdomain to another.

A range of clustering approaches for geostatistical data has been proposed over the years. They are adaptations of non-spatial clustering methods. They can be classified into four groups. The first group incorporates the spatial information by treating each observation as a point in a dimensional space formed by the geographical space and the attribute space, and a non-spatial clustering method is used subsequently. The second group uses existing non-spatial clustering methods by modifying the dissimilarity and/or similarity measure between two observations to take explicitly into account the spatial dependence [6,17,19,37]. The third group enforces the spatial contiguity during the clustering process [38,39]. The latest group relies on the assumption that observations are drawn from a particular distribution like a mixture of Gaussian or Markov random fields [1–4,14,21].

The problem of dealing with spatial correlation in spatial data mining has also been addressed in other tasks such as in predictive problems and descriptive problems. Methods to account for spatial correlation in a predictive modelling task (classification or regression) have been proposed, for instance, in references [5,16,30,43,44,50]. Approaches accounting for spatial correlation in a descriptive problem have proposed, for example, in references [29,45].

In the present paper, a spectral clustering approach designed for multivariate geostatistical data, in which spatial dependence plays an important role, is proposed. The basic idea is to include the spatial information in the clustering procedure through a nonparametric kernel estimator of the multivariate spatial dependence structure of the data. This kernel estimator is used to build a similarity measure at pairs of data locations, emphasizing the spatial correlation among data locations. The proposed spectral clustering approach is model-free, adapted to irregularly spaced data, and can produce spatially contiguous and meaningful clusters without including any geometrical constraints. It incorporates existing methods to determine the relevant number of clusters and to evaluate the contribution of variables to the clustering. The proposed spectral approach is illustrated using both multivariate synthetic and real-world datasets.

The remainder of the paper is arranged as follows: Section 2 describes the proposed spectral clustering approach through its basic ingredients. Section 3 presents a simulation study carried out to assess the performance of the proposed spectral clustering approach. Section 4 illustrates using a real-world dataset, the capability of the proposed spectral clustering method to providing spatially contiguous and meaningful clusters. Section 5 outlines concluding remarks.

2 Methodology

Consider a set of p standardized variables of interest $\{Z_1, \ldots, Z_p\}$ defined on a fixed continuous spatial domain of interest $G \subset \mathbb{R}^d, d \geq 1$ and all measured at a set of distinct locations $\{\mathbf{x}_t \in G\}_{t=1}^n$. The goal is to partition these data locations into spatially contiguous and meaningful clusters so that data locations belonging to the same cluster have a certain degree of homogeneity, while data locations in the different clusters have to be as different as possible. This section describes the different ingredients required to implement the proposed spectral clustering approach.

2.1 Similarity measure

One of the key ingredients in spectral clustering as well as in other clustering methods is the similarity measure. The traditional spectral clustering typically uses the well-known Gaussian kernel function based on the Euclidean distance in the attribute space. However, in the geostatistical framework, this type of similarity measure cannot reflect the spatial dependence structure of the data, even if geographical coordinates are also considered as attributes. The core idea is to build a similarity measure that takes into account the spatial dependency of data.

The multivariate spatial dependence structure of data is commonly described using direct and cross-variograms [47] $\{\gamma_{ij}(\mathbf{u}, \mathbf{v}) = \mathbb{C}\mathrm{ov}(Z_i(\mathbf{u}) - Z_i(\mathbf{v}), Z_j(\mathbf{u}) - Z_j(\mathbf{v}))\}_{i,j=1}^p$ defined at any pair of locations $(\mathbf{u}, \mathbf{v}) \in G^2$. Direct and cross-variograms at pair of locations can be estimated as follows:

$$\widehat{\gamma}_{ij}(\mathbf{u}, \mathbf{v}) = \frac{\sum_{l,l'=1}^n K_\lambda^\star((\mathbf{u}, \mathbf{v}), (\mathbf{x}_l, \mathbf{x}_{l'}))\, \Delta_{ij}(\mathbf{x}_l, \mathbf{x}_{l'})}{2\sum_{l,l'=1}^n K_\lambda^\star((\mathbf{u}, \mathbf{v}), (\mathbf{x}_l, \mathbf{x}_{l'}))} \mathbb{1}_{\{\mathbf{u}\neq\mathbf{v}\}} \tag{1}$$

where $\Delta_{ij}(\mathbf{x}_l, \mathbf{x}_{l'}) = (Z_i(\mathbf{x}_l) - Z_i(\mathbf{x}_{l'}))\left(Z_j(\mathbf{x}_l) - Z_j(\mathbf{x}_{l'})\right)$; $K_\lambda^\star((\mathbf{u}, \mathbf{v}), (\mathbf{x}_l, \mathbf{x}_{l'})) = K_\lambda(\|\mathbf{u} - \mathbf{x}_l\|)K_\lambda(\|\mathbf{v} - \mathbf{x}_{l'}\|)$, with $K_\lambda(\cdot)$ a nonnegative kernel function with constant bandwidth parameter $\lambda > 0$; $\mathbb{1}_{\{\mathbf{u}\neq\mathbf{v}\}}$ takes the value 1 for $\mathbf{u} \neq \mathbf{v}$ and 0 for $\mathbf{u} = \mathbf{v}$.

The nonparametric kernel estimator of the direct and cross-variograms defined in Eq. (1) has previously been employed in references [17,19]. It is the analogue of the nonparametric kernel estimator of the direct and cross-covariance functions proposed in reference [27]. It is defined at any pair of locations and not only at a pair of data locations. As highlighted in reference [27], second-order non-stationarity in data can be well captured by this type of estimator. The role of the kernel function $K_\lambda(\cdot)$ in Eq. (1) is to weight data locations according to a target location so

that data locations close to the target location receive more weight than remote data locations.

The nonparametric kernel estimator of direct and cross-variograms defined in Eq. (1) is now used to build a similarity measure that takes into account the spatial dependency of data. The similarity between two data locations $\mathbf{x}_t \in G$ and $\mathbf{x}_{t'} \in G$ $(t, t' = 1, \ldots, n)$ is defined as follows:

$$s(\mathbf{x}_t, \mathbf{x}_{t'}) = 1 - \frac{1}{\Gamma} \sum_{i,j=1}^{p} |\widehat{\gamma}_{ij}(\mathbf{x}_t, \mathbf{x}_{t'})|, \tag{2}$$

with $\Gamma = \max_{(t,t') \in \{1,\ldots,n\}^2} \sum_{i,j=1}^{p} |\widehat{\gamma}_{ij}(\mathbf{x}_t, \mathbf{x}_{t'})|$ a normalizing factor. The resulting similarity matrix at all data locations is denoted $\mathbf{S} = [s(\mathbf{x}_t, \mathbf{x}_{t'})]_{t,t'=1,\ldots,n}$. Thus, contrary to the traditional spectral clustering, here the construction of the similarity matrix takes into account the spatial dependency of the data.

In Eq. (2), the term $\frac{1}{\Gamma} \sum_{i,j=1}^{p} |\widehat{\gamma}_{ij}(\mathbf{x}_t, \mathbf{x}_{t'})|$ represents the normalized dissimilarity between data locations $\mathbf{x}_t \in G$ and $\mathbf{x}_{t'} \in G$. Thus, the dissimilarity between two data locations is defined as the normalized sum of absolute values of all direct and cross-variograms at these two data locations. The similarity measure defined in Eq. (2) satisfies requirements of a similarity measure [46]: (i) $s(\mathbf{x}_t, \mathbf{x}_{t'}) \geq 0$, (ii) $s(\mathbf{x}_t, \mathbf{x}_{t'}) = s(\mathbf{x}_{t'}, \mathbf{x}_t)$, (iii) $s(\mathbf{x}_t, \mathbf{x}_t) = 1 > 0$, and (iv) $s(\mathbf{x}_t, \mathbf{x}_{t'}) \leq 1$. It is important to highlight the use of cross-variograms instead of cross-covariance functions to describe the multivariate spatial dependence structure of data, because cross-variograms have the property of symmetry, $\gamma_{ij}(\mathbf{u}, \mathbf{v}) = \gamma_{ij}(\mathbf{v}, \mathbf{u})$. Cross-covariance functions do not satisfy the property of symmetry in general, $\big(\mathbb{C}\mathrm{ov}(Z_i(\mathbf{u}), Z_j(\mathbf{v})) \neq \mathbb{C}\mathrm{ov}(Z_i(\mathbf{v}), Z_j(\mathbf{u}))\big)$. The symmetry property is one of the requirements for a similarity measure.

As can be noted, the similarity measure defined in Eq. (2) relies on the kernel function $K_\lambda(\cdot)$ used in the estimation of the multivariate spatial dependence structure of the data (Eq. (1)). It is well known that the choice of the shape of the kernel function is less important than its bandwidth parameter [48]. The kernel function $K_\lambda(\cdot)$ is taken as the Epanechnikov kernel whose support is compact and showing optimality properties in density estimation [48]. The use of a kernel function with compact support considerably reduces the computational complexity of the similarity matrix between all data locations.

Concerning the bandwidth parameter, if its value is too small, there will not be enough data locations inside the support of the kernel function $K_\lambda(\cdot)$ to estimate the spatial dependence structure reliably. Thus, one will obtain an under smoothed estimator, with high variability. On the contrary, if the value of bandwidth parameter is too large, the resulting estimator will be over smooth and farther from the underlying spatial dependence structure of the data. An empirical rule of thumb in geostatistics is used to choose the value of the bandwidth parameter λ [12,22,24]. The bandwidth parameter λ is chosen so that the support of the kernel function $K_\lambda(\cdot)$ centred at each data location contains, at least, 35 data locations. Thus, for each data location its distance to the 35th neighbour is computed; then, the maximum of resulting distances is taken as the value of the bandwidth parameter λ. The rationale behind this choice is to have a sufficient minimum number of neighbouring data locations to estimate the spatial dependence structure reliably.

2.2 Similarity graph

In order to perform spectral clustering, data locations with pairwise similarities must be transformed into an undirected weighted graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Here, $\mathcal{V}$ is the set of vertices representing data locations. $\mathcal{E}$ is the set of edges between pairs of vertices, and each edge between two vertices v_t and $v_{t'}$ carries a nonnegative weight $w_{tt'} \geq 0$ representing strength of association between vertices. Thus, the graph $\mathcal{G}$ can be described by a nonnegative weighted n by n adjacency matrix (or affinity matrix) $\mathbf{W} = [w_{tt'}]_{t,t'=1,\ldots,n}$, where $w_{tt'}$ equals 0 if the vertices v_t and $v_{t'}$ are not connected. There are several ways to construct such an affinity matrix given a similarity measure. The most common are:

- the ε-neighbourhood graph [28,31]: any two vertices for which the similarity is greater than ε are connected. After connecting the appropriate vertices, the edge weights are assigned uniformly;
- the k-nearest neighbour graph [28,31]: two vertices v_t and $v_{t'}$ are connected if v_t is among the k most similar vertices to $v_{t'}$ or (and) vice versa. After joining the appropriate vertices, the edge weights are assigned according to the similarity measure;
- the fully connected graph [28,31]: all vertices having non-null similarities are connected each other. The edge weights are assigned according to the similarity measure.

In practice, the first two construction methods lead to a sparse graph and therefore an affinity matrix containing a high proportion of zero entries (sparse matrix). Operations on sparse matrices take up less computer memories and run faster. However, under these two construction methods, the resulting sparse affinity matrix will not reflect the spatial dependence structure of the data. Indeed, a high proportion of not necessarily zero pairwise similarities were replaced by zero, inducing a loss in the structure of the data. The third construction method is suited according to reference [31] since the similarity measure defined in Eq. (2) itself already encodes local neighbourhoods (through the kernel function $K_\lambda(\cdot)$ in Eq. (1)). Also, by connecting all data locations between them, the third construction method has the advantage to be able to produce a same cluster in different parts of the study spatial domain.

2.3 Spectral clustering algorithm

By representing data locations as a similarity graph, the clustering problem is equivalent to a graph partitioning problem, where we identify connected components with clusters. Using the normalized Laplacian matrix of the graph, the spectral clustering solution relies on the following constrained optimization problem [28,31]:

$$\min_{\mathbf{F}\in\mathbb{R}^{n\times k}} \text{Tr}\,(\mathbf{F}^T\mathbf{L}\mathbf{F}) \quad \text{subject to} \quad \mathbf{F}^T\mathbf{F}=\mathbf{I}, \tag{3}$$

where $\mathbf{F} \in \mathbb{R}^{n\times k}$ is a $(n \times k)$ real matrix consisting of orthogonal vectors; $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ is the normalized (symmetric) graph Laplacian matrix, with $\mathbf{W} = \mathbf{S}$ (the affinity matrix); $\mathbf{D}$ is a diagonal matrix whose elements are the degrees of the nodes of the graph and corresponding to $d_{tt} = \sum_{t'=1}^{n} w_{tt'}$; $\mathbf{I}$ is the identity matrix; $\text{Tr}\,(\cdot)$ denotes the trace of the matrix; and T denotes the matrix transposition.

The solution of the optimization problem defined in Eq. (3) is the matrix with the first k eigenvectors of the graph Laplacian matrix $\mathbf{L}$ arranged as columns of $\mathbf{F}$ [28,31]. For a given number of clusters k, spectral clustering algorithm finds the top k eigenvectors. These k eigenvectors define a k-dimensional projection of the data. Then, a standard clustering algorithm such as k-means is applied to the matrix whose columns are the k eigenvectors, in order to derive the final clusters of data locations. Given the classical spectral clustering algorithm [36], the proposed geostatistical spectral clustering performs the following steps:

1. Compute the similarity matrix $\mathbf{S}$ and take the affinity matrix $\mathbf{W} = \mathbf{S}$;
2. Compute the degree matrix $\mathbf{D}$;
3. Compute the graph Laplacian matrix $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$;
4. Compute the k largest eigenvalues of $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ and form the matrix $\mathbf{F} \in \mathbb{R}^{n\times k}$ whose columns are the associated k first eigenvectors of $\mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$;
5. Normalize the rows of $\mathbf{F}$ to norm 1;
6. Cluster the rows of $\mathbf{F}$ with the k-means algorithm into clusters $C_1, \ldots, C_k$;
7. Assign data location $\mathbf{x}_t$ to the same cluster the t-th row of $\mathbf{F}$ has been assigned.

2.4 Optimal number of clusters

The determination of the most appropriate number of clusters is carried out using an internal cluster validity index. Various internal cluster validity indexes have been proposed in the literature [23]. We choose the Caliński–Harabasz index [8], which relies on the between-cluster variation and within-cluster variation. Given various number of clusters $k = 2, 3, \ldots, n-1$, the optimal number of clusters is the one that maximizes the Caliński–Harabasz index:

$$CH(k) = \frac{B(k)/(k-1)}{W(k)/(n-k)}, \tag{4}$$

where $B(k) = \sum_{m=1}^{k} n_m \|\bar{\mathbf{y}}_m - \bar{\mathbf{y}}\|^2$ is the overall between-cluster variance, and $W(k) = \sum_{m=1}^{k}\sum_{t\in C_m} \|\mathbf{y}_t - \bar{\mathbf{y}}_m\|^2$ is the overall within-cluster variance; $\mathbf{y}_t \in \mathbb{R}^k$ is the vector corresponding to the t-th row of the matrix $\mathbf{F}$; $\bar{\mathbf{y}}_m = \frac{1}{n_m}\sum_{t\in C_m}\mathbf{y}_t$ is the average of points in cluster C_m; and $\bar{\mathbf{y}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t$ is the overall average; n_m is the number of points in cluster C_m.

It is important to highlight that the determination of the optimal number of clusters does not require to repeat all steps of the proposed geostatistical spectral clustering for each value of the predefined number of clusters k. Two of the three main steps (construction of the similarity matrix and computation of the eigen-decomposition of the graph Laplacian matrix) are performed only once and regardless the predefined number of clusters k. The only step which is carried out for each value of k is the k-means clustering which is quite fast.

2.5 Variable importance

After the elaboration of a clustering, it is important to know the contribution of each variable in the formation of the resulting clusters. This information can greatly improve the interpretation of these clusters. By considering variables $\{Z_1, \ldots, Z_p\}$ as predictors and cluster labels $\{C_1, \ldots, C_k\}$ as the response, the random forest classifier [7] offers a mechanism for assessing the importance of variables. Random forest consists of a number of decision trees. Every node in the decision trees is a condition on a single variable, designed to split the dataset into two so that similar response values end up in the same set. The splitting is based on a measure of impurity. For a forest, the impurity reduction from a variable $Z_j (j = 1, \ldots, p)$ can be averaged and used as a measure of importance [7]:

$$\text{Imp}(Z_j) = \frac{1}{M}\sum_{m=1}^{M}\sum_{t\in\varphi_m} \mathbb{1}_{j_t=j}\,[p(t)\Delta i(t)], \tag{5}$$

where φ_m denotes an ensemble of M randomized decision trees; j_t is the variable used at node t; $p(t)$ is the proportion n_t/n of samples reaching t and $\Delta i(t)$ is the impurity reduction at node t: $\Delta i(t) = i(t) - \frac{n_{t_L}}{n_t} i(t_L) - \frac{n_{t_R}}{n_t} i(t_R)$, where $i(t)$ is an impurity measure (e.g. Gini index) and t_L and t_R are the child nodes. $\mathbb{1}_{j_t=j}$ takes the value 1 for $j_t = j$ and 0 for $j_t \neq j$.

Equation (5) indicates the contribution of each variable to the homogeneity of the nodes and leaves in the resulting forest. Every time a split of a node is made on a specific variable; the impurity measure for the child nodes is calculated and compared to that of the parent node. Thus, the more the child nodes have lower impurity, the more the impurity reduction is higher. The changes in impurity are summed for each variable and normalized at the end of the calculation. When the Gini index is chosen as an impurity measure, the measure provided in Eq. (5) is known as the Gini importance.

3 Simulation study

A simulation study is carried out to evaluate the effectiveness of the proposed spectral clustering method to take advantage of the spatial dependence to produce spatially contiguous and meaningful clusters of data locations. Synthetic data containing known spatial clusters produced by simulation are considered. The results provided by the proposed spectral clustering method are compared with some baseline clustering methods.

3.1 Baseline clustering methods

The first baseline method (M1) is the classical k-means clustering [11]. The second baseline method (M2) is the traditional spectral clustering based on the fully connected graph [36]. The third baseline method (M3) is the spectral clustering based on k-nearest spatial neighbour graph [39]. The fourth baseline method is the spectral clustering (M4) based on the Delaunay graph [39]. In all these benchmark clustering methods, geographical coordinates are considered as attributes. For methods M2, M3, and M4, the similarity measure used for computing the graph edge weights is the Gaussian kernel function based on the Euclidean distance in the attribute space. The bandwidth parameter of the Gaussian kernel function is taken in the order of the mean distance of a point (in the attribute space) to its k-th nearest neighbour ($k \sim log(n) + 1$) as suggested in reference [31].

3.2 Data generation

Consider the bivariate Matérn stationary covariance function model defined in reference [20]:

$$C_{ii}(\mathbf{h}) = \sigma_i^2 \mathcal{M}(\mathbf{h}|\nu_i, a_i), \quad \text{for } i = 1, 2, \text{ and}$$

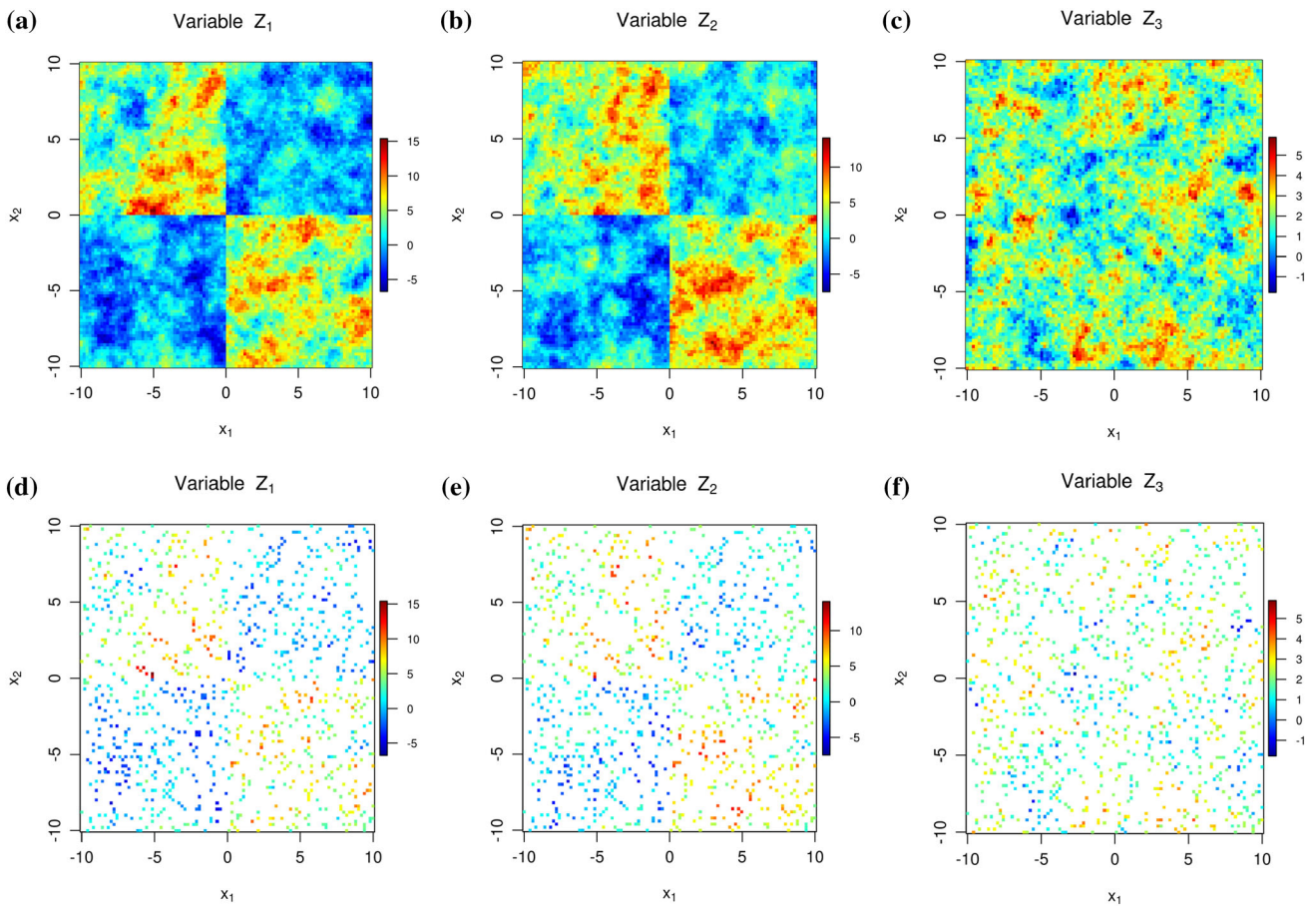

Fig. 1 **a–c** Complete data and **d–f** sampling data

$$C_{12}(\mathbf{h}) = C_{21}(\mathbf{h}) = \rho_{12}\sigma_1\sigma_2\mathcal{M}(\mathbf{h}|\nu_{12}, a_{12}), \quad (6)$$

where $\sigma_i^2 > 0$, $\nu_i > 0$, and $a_i > 0$ are, respectively, the variance parameter, the smoothness parameter, and the scale parameter; $\rho_{12}, \nu_{12}, a_{12}$ are, respectively, the co-located correlation coefficient, smoothness, and scale parameters; $\mathcal{M}(\mathbf{h}|\nu, a) = \frac{2^{1-\nu}}{\Gamma(\nu)}(a\|\mathbf{h}\|)^{\nu}\mathcal{K}_{\nu}(a\|\mathbf{h}\|)$, with $\mathcal{K}_{\nu}(\cdot)$ a modified Bessel function of the second kind of order ν.

Three variables Z_1, Z_2, and Z_3 are simulated on the spatial domain $[-10, 10]\times[-10, 10]$ as follows. On the spatial subdomains $[-10, 0] \times [-10, 0]$ and $[0, 10] \times [0, 10]$, (Z_1, Z_2) is generated according to a Gaussian stationary bivariate random function with mean vector $(0, 0)$ and bivariate spatial dependence structure given by Eq. (6) with parameters $(\sigma_1, \sigma_2, \nu_1, \nu_2, a_1, a_2, \rho_{12}, \nu_{12}, a_{12}) = (6, 6, 0.5, 0.5, 1.4, 1.4, 0.7, 0.5, 1.4)$. On the spatial subdomains $[-10, 0] \times [0, 10]$ and $[0, 10] \times [-10, 0]$ (Z_1, Z_2) is generated with respect to a Gaussian stationary bivariate random function with mean vector $(5, 5)$ and bivariate spatial dependence structure given by Eq. (6) with parameters $(\sigma_1, \sigma_2, \nu_1, \nu_2, a_1, a_2, \rho_{12}, \nu_{12}, a_{12}) = (8,8,0.5,0.5,1,1,0.7,0.5,1)$. Z_3 is simulated independently of the other two variables, and according to a Gaussian stationary univariate random function defined on the global spatial domain $[-10, 10] \times [-10, 10]$, with mean 2, variance 1, and Matérn stationary correlation function with smoothness 0.5, and scale parameter 0.6.

A representation of simulated variables over a 100×100 regular grid is given in Fig. 1a–c. As one can see, variables Z_1 and Z_2 depict two spatially contiguous clusters with relatively high and low values (Fig. 1a, b), whereas variable Z_3 does not present spatial clusters (Fig. 1c). From the realization of these variables, a dataset of 1000 observations sampled randomly is obtained as shown in Fig. 1d–f. The goal is to recover the two intrinsic spatial clusters using this dataset. We will also check whether the variable importance measure defined in Eq. (5) will give a small contribution to variable Z_3 compared to variables Z_1 and Z_2. Before performing each clustering method, all variables have been standardized. For baseline clustering methods, geographical coordinates have also been standardized.

3.3 Experimental results

The clustering results for each method are presented in Fig. 2. It appears that all baseline clustering methods (M1, M2, M3, and M4) are not able to recover the underlying spatially contiguous clusters. Under methods M1 and M2, clusters were formed by discrimination between low and high values of variables without really accounting for any spatial correlation. The failure of methods M1 and M2 for

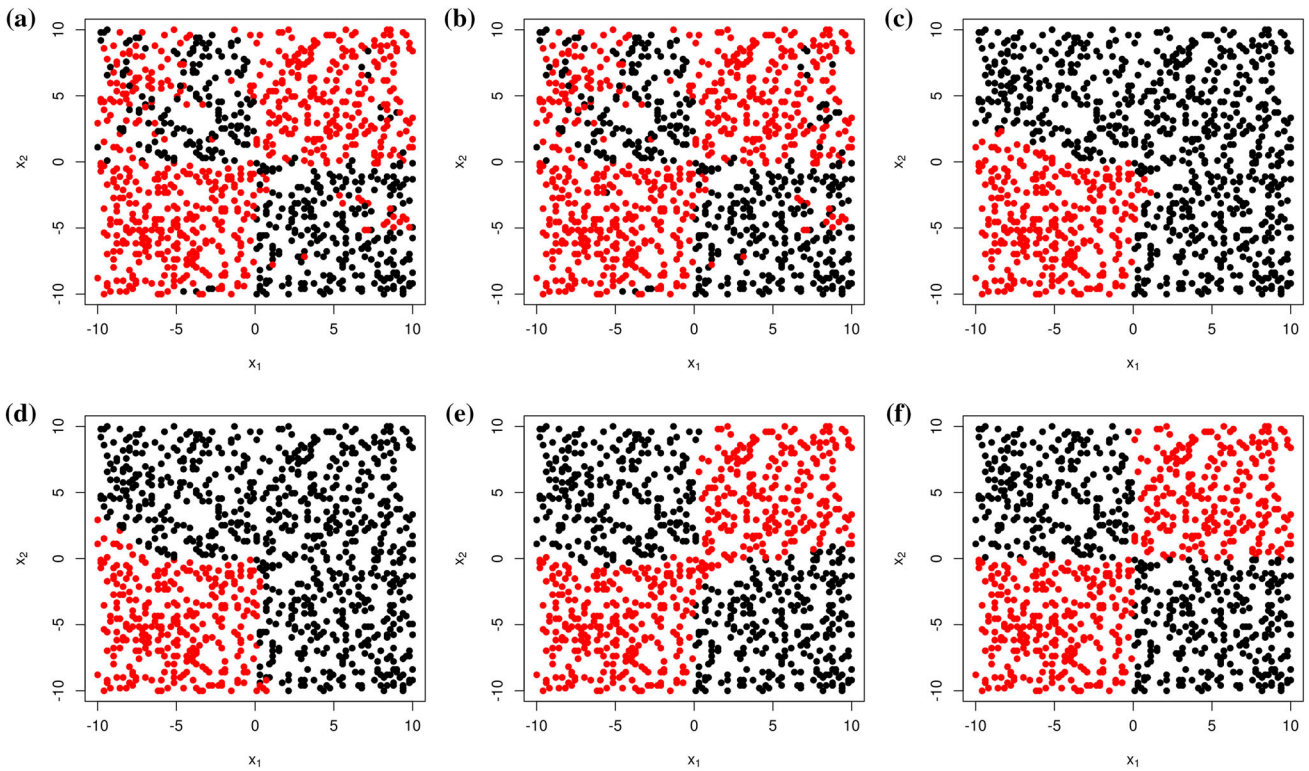

Fig. 2 **a–d** baseline clustering methods M1, M2, M3, and M4; **e** proposed spectral clustering method M5; **f** ground-truth clustering. The *colour* of *dots* identifies the cluster membership

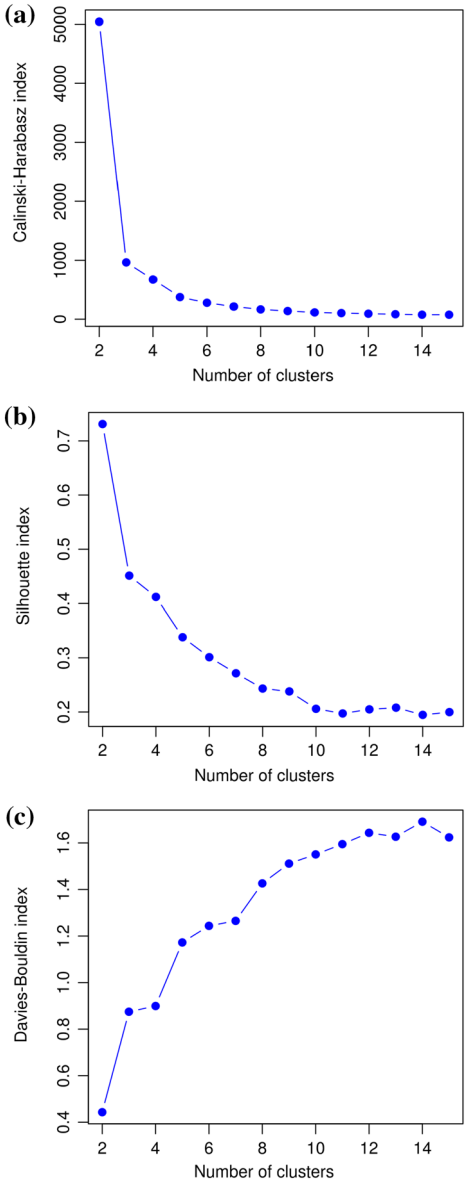

Fig. 3 Proposed spectral clustering method: selection of the optimal number of clusters through **a** Caliński–Harabasz index, **b** Silhouette index, and **c** Davies–Bouldin index

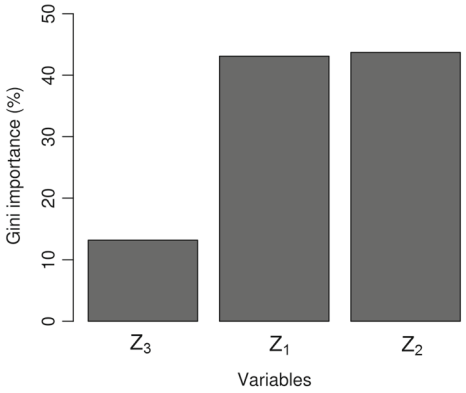

Fig. 4 Proposed spectral clustering method: contribution of each variable in the formation of the resulting clusters

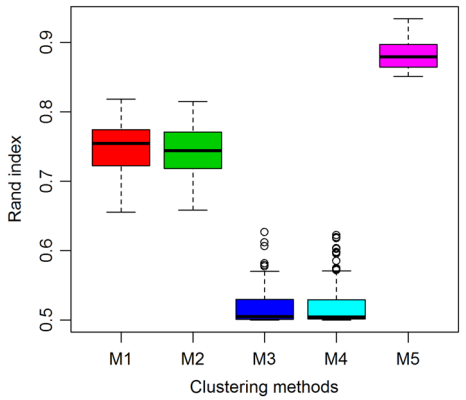

Fig. 5 Rand index between the true partition and the partition obtained for each clustering method over 100 simulated datasets

providing spatially contiguous clusters is relative to the non-distinction between the geographical space and the attribute space. Although methods M3 and M4 provide spatially contiguous clusters, they do not correspond to the underlying spatial clusters as shown in Fig. 2c, d. The inability of methods M3 and M4 for providing meaningful spatial clusters is because they are based on sparse similarity graphs which do not reflect the spatial structure of the data although they ensure the spatial contiguity. Moreover, by connecting only neighbour data locations, methods M3 and M4 are not able to detect a same cluster in different parts of the study spatial domain as is the case of this synthetic dataset. In Fig. 2e, it can be seen that the proposed spectral clustering is able to recover the two underlying spatially contiguous clusters. There are few misclassified data locations which are located at the boundaries of different spatial clusters; thereby, they are difficult to classify correctly.

Figure 3a plots the number of clusters versus the Caliński–Harabasz index defined in Eq. (4). In addition to the Caliński–Harabasz index, two other well-known internal validity indexes are plotted (Fig. 3b, c), namely silhouette index [26,40] and Davies–Bouldin index [13]. The silhouette index relies on the pairwise difference of between-cluster distances and within-cluster distances. In turn, the Davies–Bouldin index is based on a ratio of within-cluster and between-cluster distances. The maximum in the plot of the Caliński–Harabasz index (or the silhouette index) versus number of clusters is taken to indicate the underlying number of clusters. The minimum in the plot of the Davies–Bouldin index versus the number of clusters is an indication of the relevant number of clusters. As it can be noted, the Caliński–Harabasz index provides the correct number of spatial clusters as well as the two other indexes.

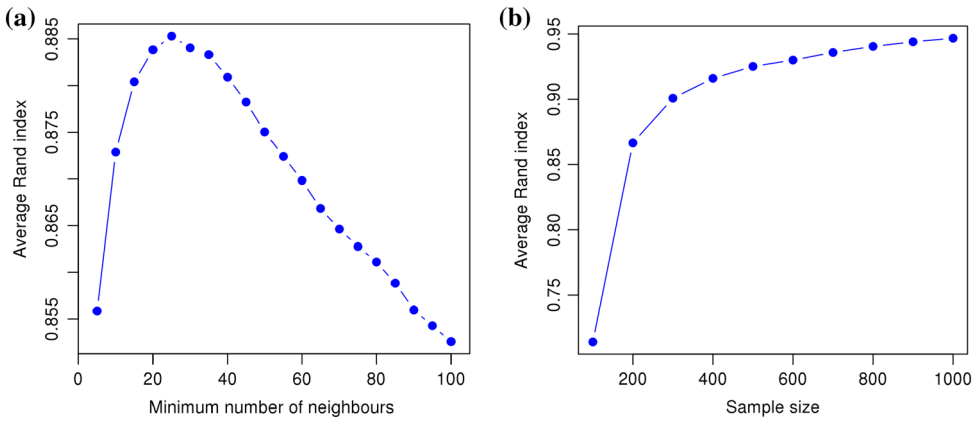

Fig. 6 Performance of the proposed clustering method under: **a** minimum numbers of neighbours and **b** sample sizes

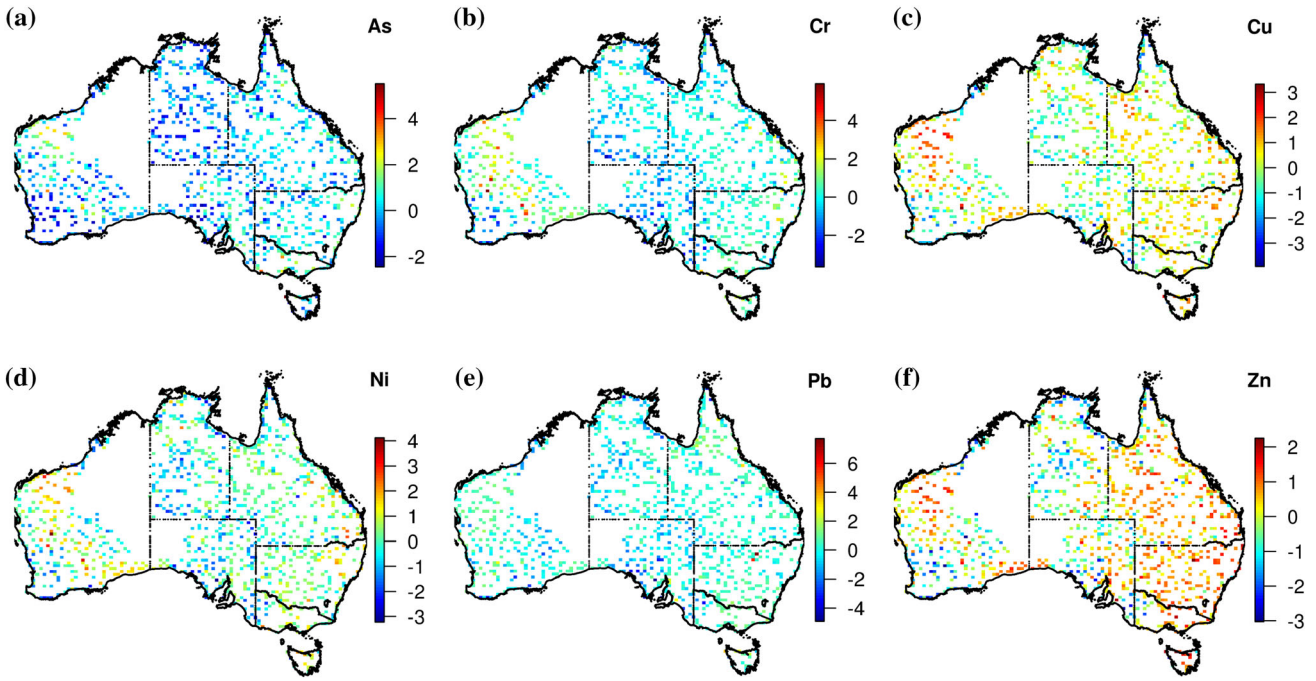

Fig. 7 Log-transformed and standardized variables for clustering purpose

As one can be noted in Fig. 2e, the presence of the non-informative variable Z_3 did not prevent the proposed spectral clustering method to retrieve the two underlying spatial clusters. The contribution of each variable in the formation of clusters is given in Fig. 4. It can be seen that the variable Z_3 has a small contribution (13%) compared to variables Z_1 and Z_2, which have relatively the same contribution (43 and 44%, respectively). Thus, the proposed spectral clustering method is robust to irrelevant variables.

The performance of the benchmark clustering methods (M1, M2, M3, and M4) and the proposed spectral clustering method are assessed using an external cluster validity index, namely the Rand index [11]. This index measures the fraction of point pairs where the resulting clustering and the ground-truth clustering agree that they belong together or do not belong together. The highest value of the statistic is one, where the clustering is perfect. Repeating independently the data generation process described in Sect. 3.2, 100 datasets of 1000 observations are formed. Each clustering method is performed to each of these 100 sampling datasets. The distribution of the Rand index computed for each clustering method is given in Fig. 5. It emerges that the baseline clustering methods (M1, M2, M3, and M4) and the proposed spectral clustering method (M5) differ notably, the latter giving the best performance. In particular, methods M1 and M2 have similar performances. Methods M3 and M4 also have similar performances, and they are the worst. These findings have been confirmed statistically by an analysis of variance with one factor (clustering method), following by the Tukey's honestly significant difference test [34].

In order to check whether the empirical rule of thumb for the selection of the bandwidth parameter turns out to be a

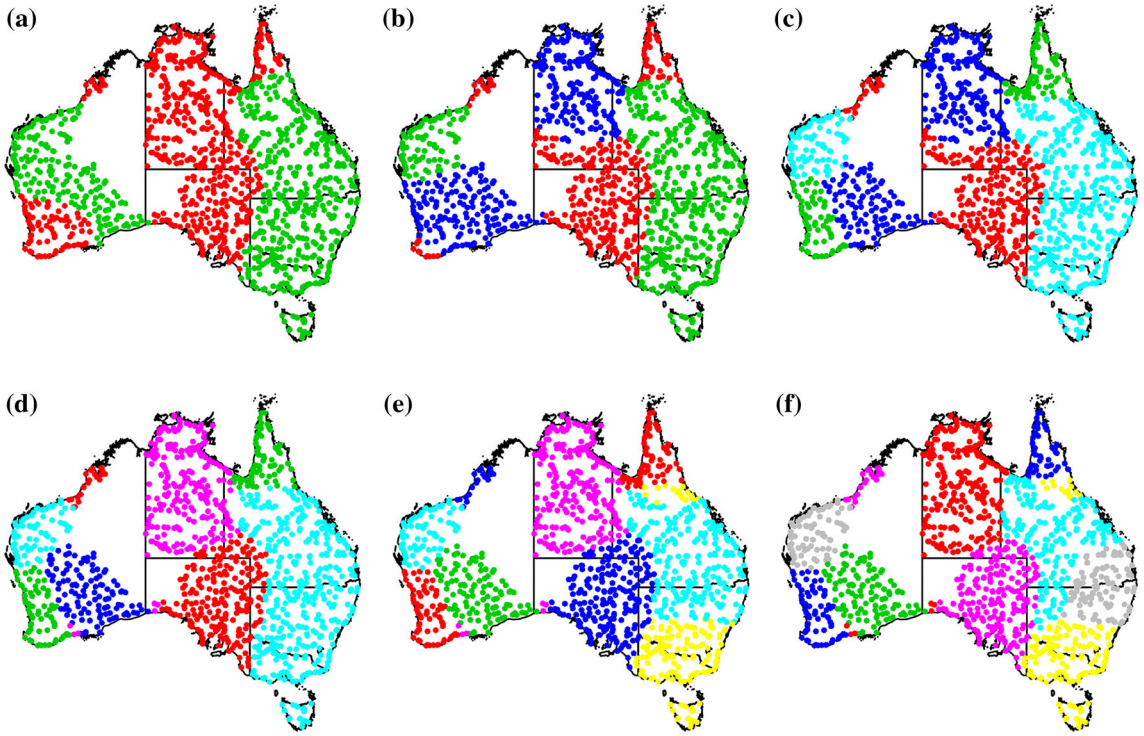

Fig. 8 Proposed spectral clustering method: **a** 2 clusters, **b** 3 clusters, **c** 4 clusters, **d** 5 clusters, **e** 6 clusters, and **f** 7 clusters. The *colour* of *dots* identifies the cluster membership

reasonable choice, a sensitivity analysis is carried out. The 100 sampling datasets obtained previously are considered. For each dataset, the proposed spectral clustering method is performed with different minimum numbers of neighbours (from 5 to 100 in steps of 5), and the Rand index of each resulting clustering is calculated. Then, for each minimum number of neighbours, the average of these 100 Rand index values is computed as presented in Fig. 6a. It appears that to take the bandwidth parameter as the maximum distance of the 35th neighbour is a reasonable choice. Indeed, as it can be seen in Fig. 6a, globally the performance of the proposed spectral clustering is highest between 20 and 35.

The performance of the proposed spectral clustering method is investigated under different sample sizes (from 100 to 1000 in steps of 100). For each sample size, 100 random samples are generated from the realization of variables Z_1, Z_2, and Z_3 presented in Fig. 1a–c; for each sample, the Rand index of the resulting clustering is computed; then, the average of these 100 Rand index values is calculated. The average Rand index for different sample sizes is given in Fig. 6b. It can be seen that globally, the performance of the proposed spectral clustering method increases with the sample size. This result makes sense because the more data we have, the more reliable the nonparametric kernel estimator of the multivariate spatial dependence structure defined in Eq. (1) will be, and so the affinity matrix.

4 Application

The proposed spectral clustering is applied to a real-world dataset from the National Geochemical Survey of Australia (NGSA) database [10]. The dataset of interest comprises six variables which are concentration elements (heavy metals) for 1314 collection sites from topsoil (0–10 cm depth) and coarse grain-size fraction (<2 mm). These six variables (concentration elements) include arsenic (As), chromium (Cr), copper (Cu), nickel (Ni), lead (Pb), and zinc (Zn). The admissible range for each variable is $[0; +\infty[$. Prior to the clustering, all variables are log-transformed because distributions of the variables are skewed and then standardized. This preliminary processing also allows to have comparable scales and identify a spatial pattern in the variables quickly. Moreover, the relative order is maintained such high transformed values correspond to high raw values and vice versa. A representation of log-transformed and standardized variables is given in Fig. 7.

Figure 8 shows the resulting spatial clusters provided by the proposed spectral clustering method for different predefined number of clusters (from 2 to 7). One can see that the proposed spectral clustering method is able to produce disconnected clusters of similar data locations. The optimal number of clusters through the Caliński–Harabasz index defined in Eq. (4) is equal to two (Fig. 9a). The plots of num-

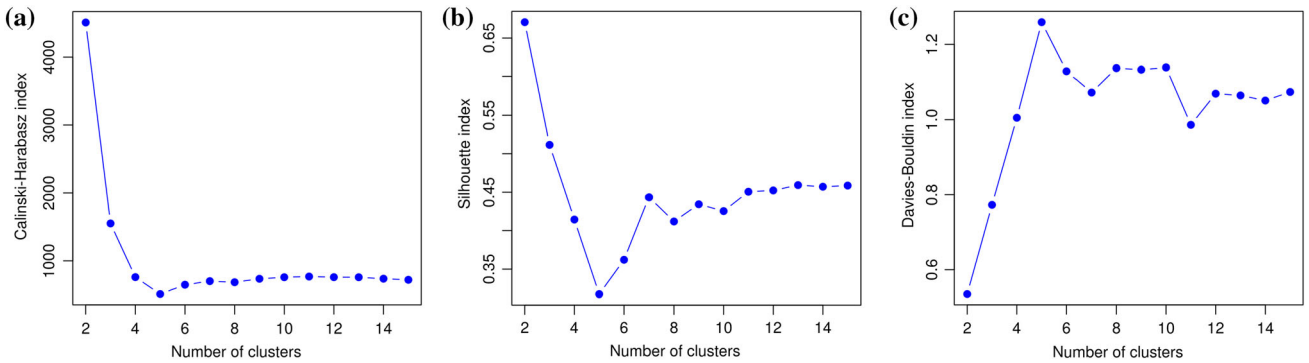

Fig. 9 Proposed spectral clustering method: **a** selection of the optimal number of clusters through **a** Caliński–Harabasz index, **b** Silhouette index, and **c** Davies–Bouldin index

Table 1 Proposed spectral clustering method: means and standard deviations of the variables (log-transformed and standardized) corresponding to the two optimal spatial clusters

	Spatial cluster 1 ($n_1 = 701$)		Spatial cluster 2 ($n_2 = 613$)	
	Mean	Std.	Mean	Std.
As	0.32	0.95	−0.36	0.93
Cr	0.42	0.90	−0.48	0.89
Cu	0.36	0.85	−0.42	0.99
Ni	0.37	0.92	−0.42	0.92
Pb	0.29	0.84	−0.33	1.06
Zn	0.43	0.79	−0.49	0.98

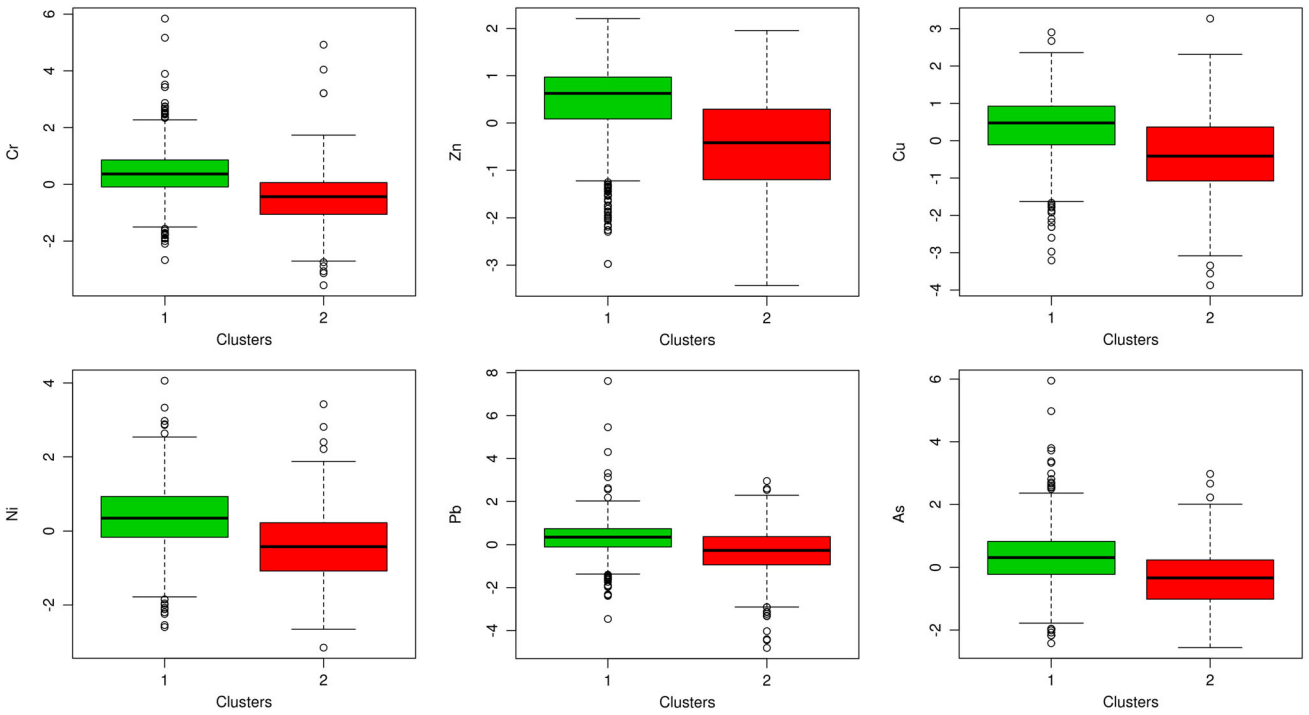

Fig. 10 Boxplot of the variables (log-transformed and standardized) corresponding to the two optimal spatial clusters

ber of clusters versus silhouette and Davies–Bouldin indexes (Fig. 9b, c) also suggest two as a suitable number of clusters.

Table 1 reports the main descriptive statistics of the variables (log-transformed and standardized) corresponding to the two optimal spatial clusters. Figure 10 depicts the boxplots of the variables corresponding to the two optimal spatial clusters. One can note that the contrast between the two spatial clusters is substantial. It appears that the spatial cluster

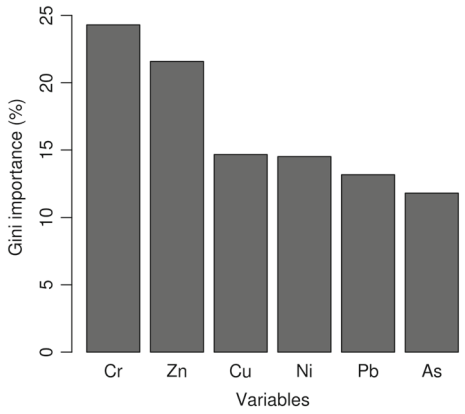

Fig. 11 Proposed spectral clustering method: contribution of each variable in the formation of the two optimal spatial clusters

1 (green points in Fig. 8a) is characterized by the highest concentrations, whereas the spatial cluster 2 shows lowest concentrations (red points in Fig. 8a). The group of lower values contains 613 observations located primarily in Northern Territory, South Australia, and South West Western Australia. The group of high values contains 701 observations located primarily in Queensland, New South Wales, Victoria, Tasmania, and a part of Western Australia.

Figure 11 shows the contribution of each variable in the formation of two optimal spatial clusters. It appears that the two most important variables are chromium (Cr) and zinc (Zn), with a relative contribution of 24 and 21%, respectively. The representation of variables such as chromium (Cr) and zinc (Zn) given in Fig. 7b, f reveals that the partition induced by the two optimal spatial clusters (Fig. 8a) is consistent with the spatial variation of these variables.

5 Concluding discussion

A spectral clustering approach aimed to discover spatially contiguous and meaningful clusters in multivariate geostatistical data has been proposed. It relies on a similarity measure built from a nonparametric kernel estimator of the multivariate spatial dependence structure of the data. As a result, it is able to produce spatially contiguous and meaningful clusters. It also incorporates existing methods to find the optimal number of clusters and to assess the contribution of variables to the clustering.

The proposed spectral clustering approach is model-free; there is no distributional assumptions or spatial dependence structure assumptions. It is adapted to irregularly spaced data and can produce spatially contiguous and meaningful clusters without including any geometrical constraints. The empirical evaluation of the proposed spectral clustering method shows that it is robust to irrelevant variables and may produce disconnected clusters of similar data locations.

The proposed spectral clustering method exploits the spatial dependence structure of the data through a nonparametric kernel estimator of this latter. Given the well-known variability of empirical estimates for small size data or sparse data, it will be difficult to estimate the multivariate spatial dependence structure reliably. In those cases, the resulting clustering could not reflect the underlying spatial clusters. When dealing with large datasets, the proposed spectral clustering method is computationally intensive. In fact, the computation of the similarity measure is more complex than calculating the sum of squared deviations, thereby increasing the overall computational complexity.

A geostatistical empirical rule of thumb has been used to choose the bandwidth parameter associated with the nonparametric kernel estimator of the multivariate spatial dependence structure of the data. Although this heuristic approach proved successful on synthetic data, it would be interesting to have an automatic bandwidth selection procedure. There exists several versions of the graph Laplacian matrix. In this paper, the normalized (symmetric) graph Laplacian matrix has been used, but a different graph Laplacian matrix may prove useful.

References

1. Allard, D.: Geostatistical classification and class kriging. J. Geogr. Inf. Decis. Anal. **2**, 87–101 (1998)
2. Allard, D., Guillot, G.: Clustering geostatistical data. In: Proceedings of the Sixth Geostatistical Conference (2000)
3. Allard, D., Monestiez, P.: Geostatistical segmentation of rainfall data. In geoENV II: Geostatistics for Environmental Applications pp. 139–150 (1999)
4. Ambroise, C., Dang, M., Govaert, G.: Clustering of spatial data by the EM algorithm. In geoENV I: Geostatistics for Environmental Applications pp. 493–504 (1995)
5. Bel, L., Allard, D., Laurent, J., Cheddadi, R., Bar-Hen, A.: CART algorithm for spatial data: application to environmental and ecological data. Comput. Stat. Data Anal. **53**, 3082–3093 (2009)
6. Bourgault, G., Marcotte, D., Legendre, P.: The multivariate (co)variogram as a spatial weighting function in classification methods. Math. Geol. **24**(5), 463–478 (1992)
7. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
8. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**(1), 1–27 (1974)
9. Cao, Y., Chen, D.R.: Consistency of regularized spectral clustering. Appl. Comput. Harmon. Anal. **30**(3), 319–336 (2011)
10. Caritat, P., Cooper, M.: National geochemical survey of Australia: The geochemical atlas of Australia. Geoscience Australia Record 2011/020 (2011)
11. Charu, C., Chandan, K.: Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, London (2013)
12. Chilès, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty. Wiley, New York (2012)
13. Davies, D.L., Bouldin, D.W.: A cluster separation measure. IEEE Trans. Pattern Anal. Mach. Intell. **1**(2), 224–227 (1979)

14. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via EM algorithm. J. R. Stat. Soc. Ser. **39**, 1–38 (1977). (with discussion)
15. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern Recogn. **41**(1), 176–190 (2008)
16. Fotheringham, A.S., Brunsdon, C., Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. Wiley, New York (2002)
17. Fouedjio, F.: A clustering approach for discovering intrinsic clusters in multivariate geostatistical data. In: Perner, P. (ed.) MLDM 2016, pp. 491–500. Springer, Berlin (2016)
18. Fouedjio, F.: Discovering spatially contiguous clusters in multivariate geostatistical data through spectral clustering. In: Li, J., et al. (eds.) ADMA 2016, pp. 547–557. Springer, Berlin (2016)
19. Fouedjio, F.: A hierarchical clustering method for multivariate geostatistical data. Spat. Stat. **18**, 334–351 (2016)
20. Gneiting, T., Kleiber, W., Schlather, M.: Cross-covariance functions for multivariate random fields. J Am. Stat. Assoc. **105**, 1167–1177 (2010)
21. Guillot, G., Kan-King-Yu, D., Michelin, J., Huet, P.: Inference of a hidden spatial tessellation from multivariate data: application to the delineation of homogeneous regions in an agricultural field. J. R. Stat. Soc. Ser. C (Appl. Stat.) **55**(3), 407–430 (2006)
22. Haas, T.C.: Lognormal and moving window methods of estimating acid deposition. J. Am. Stat. Assoc. **85**(412), 950–963 (1990)
23. Hui, X., Zhongmou, L.: Clustering validation measures. In: Charu, C., Chandan, K. (eds.) Data Clustering, pp. 571–605. Chapman and Hall/CRC, London (2013)
24. Journel, A., Huijbregts, C.: Mining Geostatistics. Blackburn Press, Caldwell (2003)
25. Kannan, R., Vempala, S., Vetta, A.: On clusterings: good, bad and spectral. J. ACM **51**(3), 497–515 (2004)
26. Kaufman, L., Rousseeuw, P.: Finding Groups in Data: An Introduction to Cluster Analysis. Wiley, New York (1990)
27. Kleiber, W., Nychka, D.: Nonstationary modeling for multivariate spatial processes. J. Multivar. Anal. **112**, 76–91 (2012)
28. Liu, J., Han, J.: Spectral clustering. In: Charu, C., Chandan, K. (eds.) Data Clustering, pp. 177–199. Chapman and Hall/CRC, London (2013)
29. Li, R., Fan, J., Jiang, J., Wu, H.: Spatiotemporal correlation in WebGIS group-user intensive access patterns. Int. J. Geogr. Inf. Sci. **31**(1), 36–55 (2017)
30. Loglisci, C., Appice, A., Malerba, D.: Collective regression for handling autocorrelation of network data in a transductive setting. J. Intell. Inf. Syst. **46**(3), 447–472 (2016)
31. Luxburg, U.V.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
32. Luxburg, U.V., Belkin, M., Bousquet, O.: Consistency of spectral clustering. Ann. Stat. **36**(2), 555–586 (2008)
33. Luxburg, U.V., Bousquet, O., Belkin, M.: Limits of spectral clustering. In: Advances in Neural Information Processing Systems. pp. 857–864 (2004)
34. Montgomery, D.: Design and Analysis of Experiments, 8th edn. Wiley, New York (2012)
35. Nascimento, M.C., de Carvalho, A.C.: Spectral methods for graph clustering a survey. Eur. J. Oper. Res. **211**(2), 221–231 (2011)
36. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Advanced in Neural Information Processing Systems. pp. 849–856. MIT Press (2001)
37. Olivier, M., Webster, R.: A geostatistical basis for spatial weighting in multivariate classification. Math. Geol. **21**, 15–35 (1989)
38. Pawitan, Y., Huang, J.: Constrained clustering of irregularly sampled spatial data. J. Stat. Comput. Simul. **73**(12), 853–865 (2003)
39. Romary, T., Ors, F., Rivoirard, J., Deraisme, J.: Unsupervised classification of multivariate geostatistical data: two algorithms. Comput. Geosci. **85**(Part B), 96–103 (2015)
40. Rousseeuw, P.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)
41. Schaeffer, S.E.: Graph clustering. Comput. Sci. Rev. **1**(1), 27–64 (2007)
42. Schuenemeyer, J., Drew, L.: Statistics for Earth and Environmental Scientists. Wiley, New York (2011)
43. Stojanova, D., Ceci, M., Appice, A., Džeroski, S.: Network regression with predictive clustering trees. Data Min. Knowl. Discovery **25**(2), 378–413 (2012)
44. Stojanova, D., Ceci, M., Appice, A., Malerba, D., Džeroski, S.: Dealing with spatial autocorrelation when learning predictive clustering trees. Ecol. Inf. **13**, 22–39 (2013)
45. Tao, J., Chloissnig, S., Karl, W.: Analysis of the spatial and temporal locality in data accesses. In: Computational Science – ICCS 2006: 6th International Conference, Reading, UK, May 28–31, 2006. Proceedings, Part II. Springer. pp. 502–509 (2006)
46. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, London (2009)
47. Wackernagel, H.: Multivariate Geostatistics: An Introduction with Applications. Springer, Berlin (2003)
48. Wand, M., Jones, C.: Kernel Smoothing. Monographs on Statistics and Applied Probability. Chapman and Hall, London (1995)
49. Zha, H., He, X., Ding, C., Gu, M., Simon, H.D.: Spectral relaxation for k-means clustering. In: Advances in neural information processing systems. pp. 1057–1064 (2001)
50. Zhao, M., X. Li, X.: An application of spatial decision tree for classification of air pollution index. In: 19th International Conference on Geoinformatics. IEEE Computer Society. pp. 1–6 (2011)