Discovering Spatially Contiguous Clusters in Multivariate Geostatistical Data Through Spectral Clustering

Francky Fouedjio(✉)

CSIRO Mineral Resources, Perth, WA, Australia
francky.fouedjiokameni@csiro.au

Abstract. Spectral clustering has recently become one of the most popular modern clustering algorithms for traditional data. However, the application of this clustering method on geostatistical data produces spatially scattered clusters, which is undesirable for many geoscience applications. In this work, we develop a spectral clustering method aimed to discover spatially contiguous and meaningful clusters in multivariate geostatistical data, in which spatial dependence plays an important role. The proposed spectral clustering method relies on a similarity measure built from a non-parametric kernel estimator of the multivariate spatial dependence structure of the data, emphasizing the spatial correlation among data locations. The capability of the proposed spectral clustering method to provide spatially contiguous and meaningful clusters is illustrated using the European Geological Surveys Geochemical database.

Keywords: Geostatistics · Spectral clustering · Spatial dependency · Spatial contiguity

1 Introduction

In recent years, spectral clustering has become one of the most popular modern clustering algorithms for classical data [11,19,22,23,27]. Spectral clustering is a class of partitional clustering algorithms that relies on the eigendecomposition of feature similarity matrices to partition the data points. 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 algorithms such as K-means algorithm, and its well-established theoretical properties [7,17,20,21,31].

However, applied to geostatistical data, spectral clustering method tends to produce spatially scattered clusters, which undesirable for many geoscience applications. This clustering method can not produce spatially contiguous and meaningful clusters because it makes the assumption that observations are independent. This fundamental assumption, however, does not hold in the realm of spatial data. Geostatistical data distinguish themselves from conventional data in that they often show properties of spatial dependency and heterogeneity over

J. Li et al. (Eds.): ADMA 2016, LNAI 10086, pp. 547–557, 2016.
DOI: 10.1007/978-3-319-49586-6_38

the study domain. Observations located close to one another in the geographical space might have similar characteristics. Furthermore, the mean, the variance, and the spatial dependence structure can be different from one sub-domain to another.

Existing clustering approaches which take into account the specificities of geostatistical data can be classified into four groups: (1) non-spatial clustering with geographical coordinates as additional variables, (2) non-spatial clustering based on a spatial dissimilarity measure, (3) spatially constrained clustering, and (4) model-based clustering. 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, for a non-spatial clustering method. The second group uses existing non-spatial clustering methods by modifying the dissimilarity measure between two observations to take explicitly into account the spatial dependence [5,12,13,24]. The third group considers spatial contiguity constraints (rather than spatial dissimilarities) in a clustering process [25,26]. The latest group is not model-free. It relies on the assumption that observations are drawn from a particular distribution like a mixture of Gaussian or Markov random fields [1–4,10,14].

In this work, we propose a spectral clustering method designed for multivariate geostatistical data, in which spatial dependence plays an important role. The basic idea is to include the spatial information in the clustering procedure through a non-parametric kernel estimator of the multivariate spatial dependence structure of the data. This estimator is used to build a measure of similarity between two data locations, emphasizing the spatial correlation among data locations. The proposed clustering method is non-parametric, adapted to irregularly spaced data, and can produce spatially contiguous clusters without including any geometrical constraints. The proposed spectral clustering method is illustrated using the European Geological Surveys Geochemical database. The results derived from the proposed spectral clustering method are compared with those provided by two baseline clustering methods: K-means clustering and traditional spectral clustering.

The rest of the paper is organized as follows. Section 2 describes the proposed spectral clustering method through its basic ingredients. Section 3 illustrates using the European Geological Surveys Geochemical database, the capability of the proposed clustering method to providing spatially contiguous and meaningful clusters. Section 4 outlines concluding remarks.

2 Method

We consider a set of p standardized variables of interest $\{Z_1, \ldots, Z_p\}$ defined on a continuous domain of interest $G \subset \mathbb{R}^d (d \geq 1)$, and all measured at a set of distinct locations $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The goal is to partition these data locations into spatially contiguous and meaningful clusters so that data locations belonging to the same cluster are more similar than those in different clusters. We describe in this section the different ingredients required to implement the proposed spectral clustering method.

2.1 Similarity Measure

One of the key tasks in spectral clustering as well as in other clustering methods is the choice of the similarity measure. The traditional spectral clustering usually calculates the similarity using the well-known Gaussian kernel based on the Euclidean distance in the attribute space. However, in the geostatistical setting, this type of similarity measure can not reflect the spatial dependence structure of the data, even if geographical coordinates are also considered as attributes. We propose a novel similarity measure that takes care of the spatial dependence between observations.

A non-parametric kernel estimator of the multivariate spatial dependence structure of the data described by the direct and cross variograms, at two locations $\mathbf{u} \in G$ and $\mathbf{v} \in G$ is given by:

$$\widehat{\gamma}_{ij}(\mathbf{u},\mathbf{v}) = \frac{\sum_{l,l'=1}^{n} K_{\epsilon}^{\star}\left((\mathbf{u},\mathbf{v}),(\mathbf{x}_l,\mathbf{x}_{l'})\right)\left(Z_i(\mathbf{x}_l) - Z_i(\mathbf{x}_{l'})\right)\left(Z_j(\mathbf{x}_l) - Z_j(\mathbf{x}_{l'})\right)}{2\sum_{l,l'=1}^{n} K_{\epsilon}^{\star}\left((\mathbf{u},\mathbf{v}),(\mathbf{x}_l,\mathbf{x}_{l'})\right)} \mathbb{1}_{\{\mathbf{u}\neq\mathbf{v}\}}, \tag{1}$$

where $(i,j) \in \{1,\ldots,p\}^2$; $K_{\epsilon}^{\star}\left((\mathbf{u},\mathbf{v}),(\mathbf{x}_l,\mathbf{x}_{l'})\right) = K_{\epsilon}(\|\mathbf{u}-\mathbf{x}_l\|)K_{\epsilon}(\|\mathbf{v}-\mathbf{x}_{l'}\|)$, with $K_{\epsilon}(\cdot)$ a non-negative kernel function with constant bandwidth parameter $\epsilon > 0$; $\mathbb{1}$ denotes the indicator function.

Given the set of estimated direct and cross variograms $\{\widehat{\gamma}_{ij}(\cdot,\cdot)\}_{i,j=1}^{p}$, the similarity between two sample locations $\mathbf{x}_t$ and $\mathbf{x}_{t'}$ $(t, t' = 1,\ldots,n)$ is defined by:

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

with$\Gamma = \max_{(t,t')\in\{1,\ldots,n\}^2}\sum_{i,j=1}^{p}|\widehat{\gamma}_{ij}(\mathbf{x}_t,\mathbf{x}_{t'})|$. The resulting similarity matrix at all data locations is denoted $\mathbf{S} = [s(\mathbf{x}_t,\mathbf{x}_{t'})]_{t,t'=1,\ldots,n}$.

In Eq. (2), the term $\frac{1}{\Gamma}\sum_{i,j=1}^{p}|\widehat{\gamma}_{ij}(\mathbf{x}_t,\mathbf{x}_{t'})|$ represents the dissimilarity (normalized) between data locations $\mathbf{x}_t$ and $\mathbf{x}_{t'}$. Thus, the dissimilarity between two data locations is defined as the sum (normalized) of absolute values of all direct and cross variograms at these two data locations. Equation (2) well defines a measure of similarity [28].

2.2 Similarity Graph

Spectral clustering requires that the data are represented in the form of an undirected similarity graph $\mathcal{G} = (\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ is the set of edges between pairs of vertices. We construct a graph $\mathcal{G}$ from the similarity measure defined in Eq. (2), where the vertices of the graph represent the data locations, and the edge weights represent similarities between data locations. The similarity graph resulting from this construction is a full connected (complete) graph. This construction is suited according to [19] since the similarity measure defined in Eq. (2) itself already encodes local neighbourhoods (through the kernel function $K_\epsilon(\cdot)$ in Eq. (1)). Moreover, this construction is coherent with the Tobler's first law of geography [29]: everything is related to everything else, but near things are more related than distant things.

Given the similarity graph, the next step in the spectral clustering is the computation of the graph Laplacian matrix. There are several versions of the graph Laplacian matrix [19,20]. We use the normalized graph Laplacian matrix defined as: $\mathbf{L} = \mathbf{D}^{-1/2}(\mathbf{D}-\mathbf{S})\mathbf{D}^{-1/2} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$, where $\mathbf{S}$ is the affinity (similarity) matrix between every pair of the data locations built from the similarity measure defined in Eq. (2); $\mathbf{D}$ is a diagonal matrix whose elements are the degrees of the nodes of the graph $\mathcal{G}$ and corresponding to $d_{tt} = \sum_{t'=1}^{n} s(\mathbf{x}_t, \mathbf{x}_{t'})$; $\mathbf{I}$ denotes the identity matrix.

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. For a given number of clusters q, spectral clustering algorithm finds the top q eigenvectors. These q eigenvectors define a q-dimensional projection of the data. Then, a standard clustering algorithm such as K-means is applied to derive the final clusters of the data locations. The proposed spectral clustering algorithm for multivariate geostatistical data performs the following steps:

1. compute the similarity matrix of all data locations $\mathbf{S}$;
2. compute the degree matrix $\mathbf{D}$;
3. compute the graph Laplacian matrix $\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$;
4. compute the q largest eigenvalues of $\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$ and form the matrix $\mathbf{F} \in \mathbb{R}^{n\times q}$ whose columns are the associated q first eigenvectors of $\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$;
5. normalize the rows of $\mathbf{F}$ to norm 1;
6. cluster the rows of $\mathbf{F}$ with the K-means algorithm into clusters $C_1, \ldots, C_q$;
7. assign data location $\mathbf{x}_t$ to the same cluster the row t of $\mathbf{F}$ has been assigned.

2.4 Hyper-parameters Selection

The proposed spectral clustering method relies on the kernel function $K_\epsilon(\cdot)$ used in the estimation of the multivariate spatial dependence structure of the data (Eq. (1)). The choice of the kernel function $K_\epsilon(\cdot)$ is less important than the choice of its bandwidth parameter ϵ. We opt for the Epanechnikov kernel whose support is compact, showing optimality properties in density estimation [30]. To estimate the spatial dependence structure of the data reliably, the bandwidth parameter ϵ is chosen by using an empirical rule of thumb in geostatistics [9,15,16]: ϵ is chosen so that the support of the kernel function $K_\epsilon(\cdot)$ centered at each data location contains at least 35 observations. 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 ϵ.

The optimal number of clusters is chosen so that it corresponds to the best clustering identified in terms of an internal clustering validation measure. A variety of internal cluster validation indexes has been proposed in the literature [8,28]. We choose the Caliński-Harabasz index [6] which is best suited for K-means clustering solutions with squared Euclidean distances. Given various number of clusters $q = 2, 3, \ldots$, the optimal number of clusters is the one that maximizes the Caliński-Harabasz index:

$$CH(q) = \frac{B(q)/(q-1)}{W(q)/(n-q)}, \tag{3}$$

where $B(q) = \sum_{m=1}^{q} n_m \|\bar{\mathbf{y}}_m - \bar{\mathbf{y}}\|^2$ is the overall between-cluster variance, and $W(q) = \sum_{m=1}^{q} \sum_{t \in C_m} \|\mathbf{y}_t - \bar{\mathbf{y}}_m\|^2$ is the overall within-cluster variance; $\mathbf{y}_t \in \mathbb{R}^q$ 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.

3 Application

The proposed spectral clustering method is applied to the European Geological Surveys Geochemical database. The results provided by the proposed spectral clustering method are compared with those produced by two baseline clustering methods: K-means clustering and traditional spectral clustering.

3.1 Dataset

Data correspond to eight critical heavy metals in topsoils from the European Geological Surveys Geochemical database (26 European countries) [18]. Variables are: arsenic (As), cadmium (Cd), chromium (Cr), copper (Cu), mercury (Hg), nickel (Ni), lead (Pb), and zinc (Zn). On 1588 georeferenced available data, 1498 observations have been used in this application because there are some missing values for some variables. Prior to the clustering, all variables are logit-transformed and standardized. A representation of logit-transformed and standardized variables is given in Fig. 1. In the two baseline clustering methods, geographical coordinates are considered as attributes.

3.2 Results

Figure 2 shows the results provided by the baseline clustering methods and the proposed spectral clustering method, for different predefined number of clusters (from 2 to 4). As one can see, the baseline clustering (non-spatial clustering) methods fail to produce spatially contiguous clusters. The failure of these clustering methods is not surprising because they do not distinguish between the geographical space and the attribute space. It appears that the proposed spectral clustering method can produce spatially contiguous clusters. Moreover, the

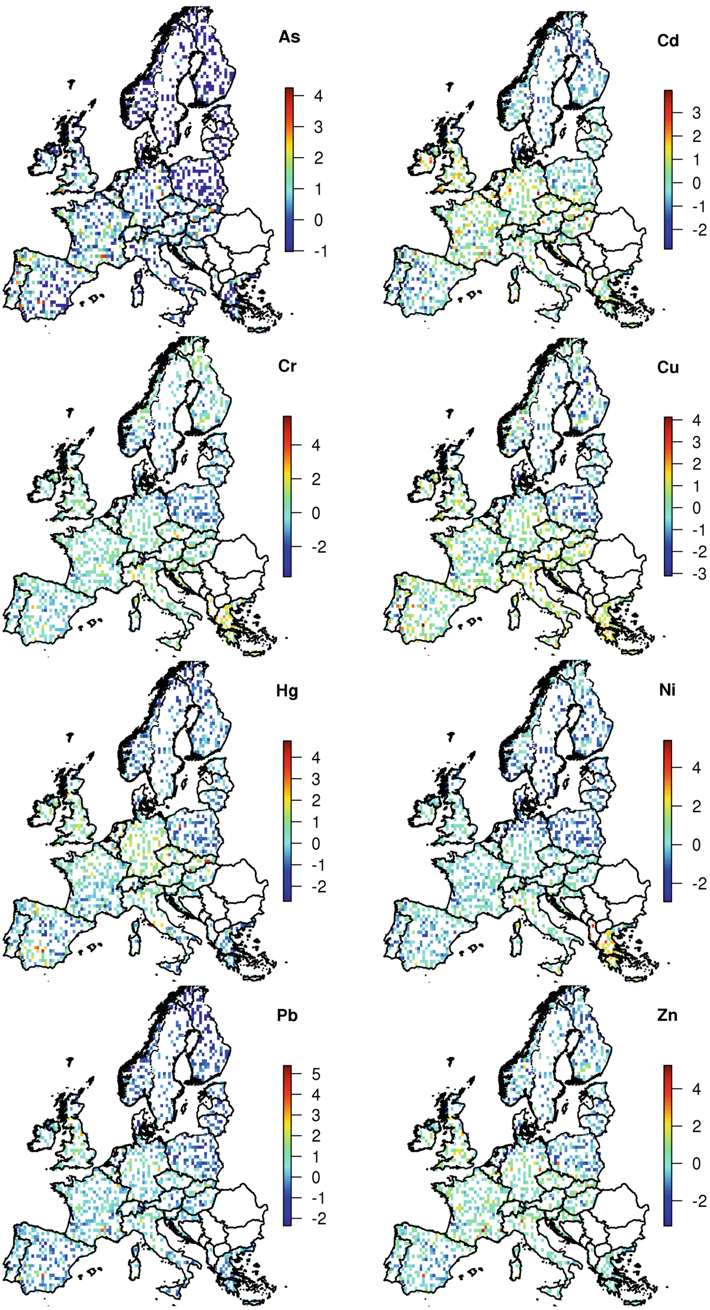

Fig. 1. Logit-transformed and standardized variables for clustering purpose. (Color figure online)

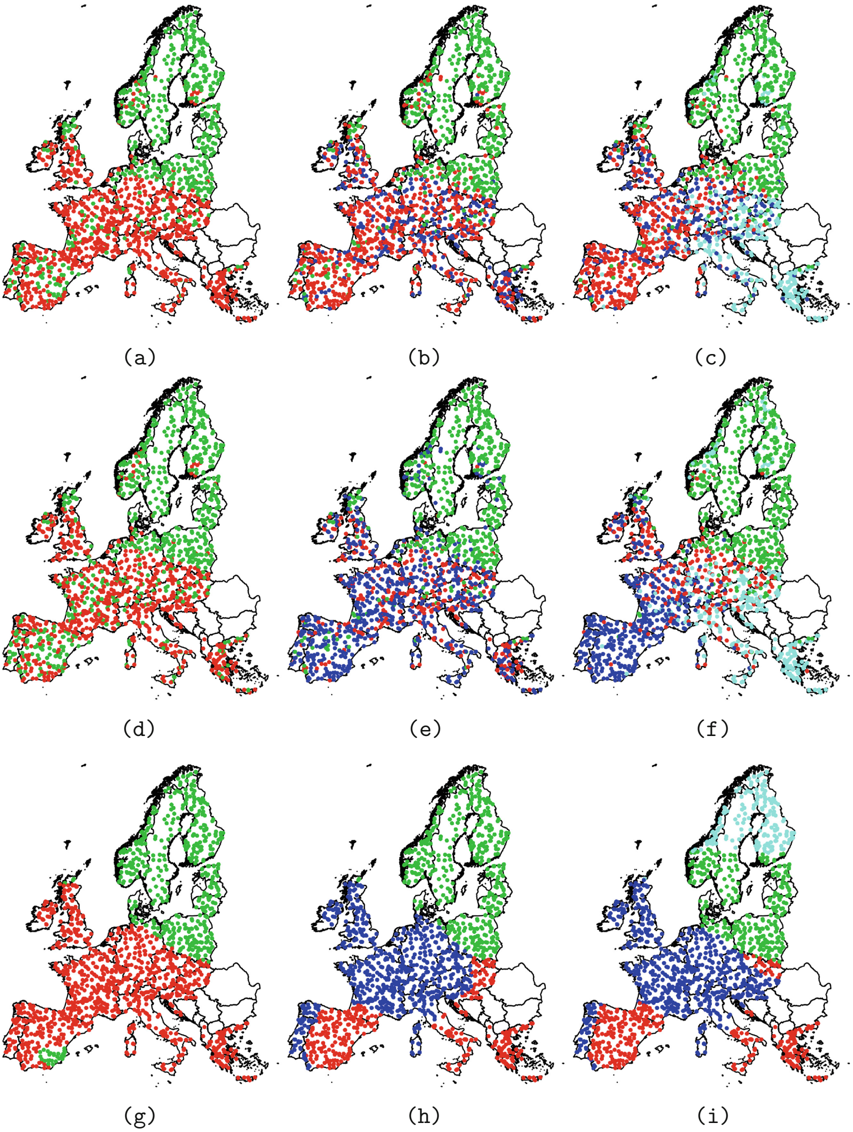

Fig. 2. (a, b, c) K-means clustering for 2, 3, and 4 clusters; (d, e, f) Traditional spectral clustering for 2, 3, and 4 clusters; (g, h, i) Proposed spectral clustering for 2, 3, and 4 clusters. The color of dots identifies the cluster membership. (Color figure online)

proposed spectral clustering method can produce disconnected clusters of similar data locations.

In the proposed spectral clustering method, the optimal number of clusters through the Caliński-Harabasz index defined in Eq. (3) corresponds to two as shown in Fig. 3. Table 1 reports the means and standard deviations of the variables (Logit-transformed and standardized) corresponding to the two optimal spatial clusters. There is a marked difference between the properties of samples in each spatial cluster. It appears that spatial cluster 1 (green points in Fig. 2g) is characterized by the lowest concentrations; whereas spatial cluster 2 shows highest concentrations (red points in Fig. 2g). The group of lower values contains 494 observations located primarily in countries of Northern Europe (Denmark, Norway, Sweden, Finland, Estonia, Latvia, and Lithuania). The group of high values contains 1004 observations located in United Kingdom, Ireland, countries of Western Europe, and countries of Southern Europe.

After the elaboration of a clustering, it is important to know the contribution of each variable in the formation of the resulting clusters. By considering variables as predictors and cluster labels as the response, the random forest classifier is used to provide the importance of variables as shown in Fig. 4. It appears that the two most important variables are arsenic (As) and lead (Pb), with a relative contribution of 19 % and 18 % respectively. This result is explained by the fact that the contrast between spatial clusters 1 and 2 is more pronounced for these two variables compared to other variables as one can see in Table 1. Moreover, a visual inspection of the variables arsenic (As) and lead (Pb) (Fig. 1) shows that the partition given by spatial clusters 1 and 2 (Fig. 2g) is coherent with the spatial variation of these variables.

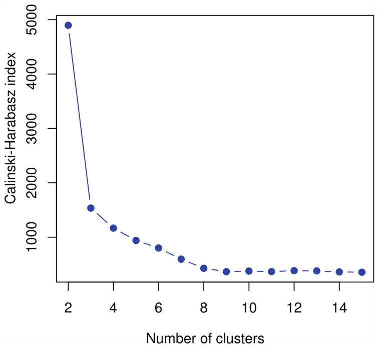

Fig. 3. Proposed spectral clustering method: selection of the optimal number of clusters through CH index.

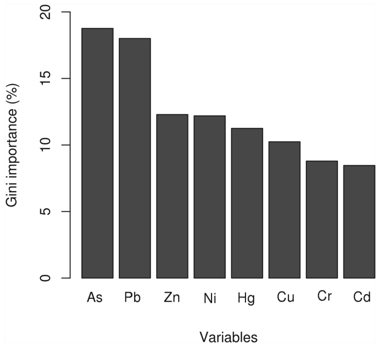

Fig. 4. Proposed spectral clustering method: contribution of each variable in the formation of the two optimal spatial clusters based on the Gini importance measure of the random forest classifier.

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

	Spatial cluster 1 ($n_1 = 494$)		Spatial cluster 2 ($n_2 = 1004$)	
	Mean	Std.	Mean	Std.
As	-0.77	0.54	0.38	0.95
Cd	-0.50	0.79	0.25	1.00
Cr	-0.49	0.95	0.24	0.93
Cu	-0.65	0.94	0.32	0.86
Hg	-0.60	0.81	0.29	0.95
Ni	-0.65	0.77	0.32	0.94
Pb	-0.72	0.91	0.36	0.84
Zn	-0.62	0.88	0.31	0.91

4 Conclusion

In this work, a spectral clustering method aimed to discover spatially contiguous and meaningful clusters in multivariate geostatistical data has been developed. The proposed spectral clustering method relies on a similarity measure built from a non-parametric kernel estimator of the multivariate spatial dependence structure of the data, thereby reinforcing the spatial contiguity of the resulting clusters. The proposed spectral clustering approach is non-parametric; there is no distributional assumptions or spatial dependence structure assumptions. It is adapted to irregularly sampled data and can produce spatially contiguous and meaningful clusters without including any geometrical constraints. Applied to the European Geological Surveys Geochemical database, the proposed spectral clustering method highlights two spatially contiguous clusters with significant meaning. It is also able to produce disconnected clusters of similar data locations. The proposed spectral clustering method is computationally intensive when dealing with large datasets. Indeed, the calculation of the similarity matrix at all data locations is more complex than calculating the sum of squared deviations. Future work includes the application of the proposed spectral clustering method to other geostatistical databases.

References

1. Allard, D.: Geostatistical classification and class kriging. J. Geog. 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. 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)
6. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Commun. Stat. **3**(1), 1–27 (1974)
7. Cao, Y., Chen, D.R.: Consistency of regularized spectral clustering. Appl. Comput. Harmonic Anal. **30**(3), 319–336 (2011)
8. Charu, C., Chandan, K.: Data Clustering: Algorithms and Applications. Chapman and Hall/CRC, Boca Raton (2013)
9. Chilès, J.P., Delfiner, P.: Geostatistics: Modeling Spatial Uncertainty. Wiley, Hoboken (2012)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via EM algorithm (with discussion). J. Roy. Stat. Soc. Ser. **39**, 1–38 (1977)
11. Filippone, M., Camastra, F., Masulli, F., Rovetta, S.: A survey of kernel and spectral methods for clustering. Pattern Recogn. **41**(1), 176–190 (2008)
12. Fouedjio, F.: A clustering approach for discovering intrinsic clusters in multivariate geostatistical data. In: Perner, P. (ed.) MLDM 2016. LNCS, vol. 9729, pp. 491–500. Springer, Switzerland (2016)
13. Fouedjio, F.: A hierarchical clustering method for multivariate geostatistical data. Spatial Statistics (2016)
14. 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. Roy. Stat. Soc. Ser. C (Appl. Stat.) **55**(3), 407–430 (2006)
15. Haas, T.C.: Lognormal and moving window methods of estimating acid deposition. J. Am. Stat. Assoc. **85**(412), 950–963 (1990)
16. Journel, A., Huijbregts, C.: Mining Geostatistics. Blackburn Press, New York (2003)
17. Kannan, R., Vempala, S., Vetta, A.: On clusterings: good, bad and spectral. J. ACM **51**(3), 497–515 (2004)

18. Lado, L., Hengl, T., Reuter, I.: Heavy metals in European soils: a geostatistical analysis of the FOREGS geochemical database. Geoderma **148**(2), 189–199 (2008)
19. Luxburg, U.V.: A tutorial on spectral clustering. Stat. Comput. **17**(4), 395–416 (2007)
20. Luxburg, U.V., Belkin, M., Bousquet, O.: Consistency of spectral clustering. Ann. Stat. **36**(2), 555–586 (2008)
21. Luxburg, U.V., Bousquet, O., Belkin, M.: Limits of spectral clustering. In: Advances in Neural Information Processing Systems, pp. 857–864 (2004)
22. Nascimento, M.C., Carvalho, A.C.: Spectral methods for graph clustering – a survey. Eu. J. Oper. Res. **211**(2), 221–231 (2011)
23. Ng, A.Y., Jordan, M.I., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems, pp. 849–856. MIT Press (2001)
24. Olivier, M., Webster, R.: A geostatistical basis for spatial weighting in multivariate classification. Math. Geol. **21**, 15–35 (1989)
25. Pawitan, Y., Huang, J.: Constrained clustering of irregularly sampled spatial data. J. Stat. Comput. Simul. **73**(12), 853–865 (2003)
26. Romary, T., Ors, F., Rivoirard, J., Deraisme, J.: Unsupervised classification of multivariate geostatistical data: two algorithms. Comput. Geosci. **85**, 96–103 (2015)
27. Schaeffer, S.E.: Graph clustering. Comput. Sci. Rev. **1**(1), 27–64 (2007)
28. Theodoridis, S., Koutroumbas, K.: Pattern Recognition, 4th edn. Academic Press, New York (2009)
29. Tobler, W.R.: A computer movie simulating urban growth in the Detroit region. Econ. Geogr. **46**, 234–240 (1970)
30. Wand, M., Jones, C.: Kernel Smoothing. Monographs on Statistics and Applied Probability. Chapman & Hall, Sanford (1995)
31. 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)