

2020

## Towards geostatistical learning for the geosciences: A case study in improving the spatial awareness of spectral clustering

Hassan Talebi  
*Edith Cowan University*

L. J. M. Peeters

Ute Mueller  
*Edith Cowan University*

R. Tolosana-Delgado

K. G. van den Boogaart

Follow this and additional works at: <https://ro.ecu.edu.au/ecuworkspost2013>



Part of the [Earth Sciences Commons](#)

---

[10.1007/s11004-020-09867-0](https://doi.org/10.1007/s11004-020-09867-0)

Talebi, H., Peeters, L. J. M., Mueller, U., Tolosana-Delgado, R., & van den Boogaart, K. G. (2020). Towards geostatistical learning for the geosciences: A case study in improving the spatial awareness of spectral clustering. *Mathematical Geosciences*, 52(8), 1035-1048. <https://doi.org/10.1007/s11004-020-09867-0>

This Journal Article is posted at Research Online.  
<https://ro.ecu.edu.au/ecuworkspost2013/10813>



# Towards Geostatistical Learning for the Geosciences: A Case Study in Improving the Spatial Awareness of Spectral Clustering

H. Talebi<sup>1,2</sup> · L. J. M. Peeters<sup>1</sup> · U. Mueller<sup>2</sup> ·  
R. Tolosana-Delgado<sup>3</sup> · K. G. van den Boogaart<sup>3</sup>

Received: 19 December 2019 / Accepted: 7 April 2020 / Published online: 8 June 2020  
© The Author(s) 2020

**Abstract** The particularities of geosystems and geoscience data must be understood before any development or implementation of statistical learning algorithms. Without such knowledge, the predictions and inferences may not be accurate and physically consistent. Accuracy, transparency and interpretability, credibility, and physical realism are minimum criteria for statistical learning algorithms when applied to the geosciences. This study briefly reviews several characteristics of geoscience data and challenges for novel statistical learning algorithms. A novel spatial spectral clustering approach is introduced to illustrate how statistical learners can be adapted for modelling geoscience data. The spatial awareness and physical realism of the spectral clustering are improved by utilising a dissimilarity matrix based on nonparametric higher-order spatial statistics. The proposed model-free technique can identify meaningful spatial clusters (i.e. meaningful geographical subregions) from multivariate spatial data at different scales without the need to define a model of co-dependence. Several mixed (e.g. continuous and categorical) variables can be used as inputs to the proposed clustering technique. The proposed technique is illustrated using synthetic and real mining datasets. The results of the case studies confirm the usefulness of the proposed method for modelling spatial data.

**Keywords** Statistical learning · Spatial data · Spatial clustering · Higher-order spatial statistics · Geoscience data

---

✉ H. Talebi  
Hassan.Talebi@csiro.au

<sup>1</sup> CSIRO Deep Earth Imaging FSP, Perth, Australia

<sup>2</sup> School of Science, Edith Cowan University, Joondalup, Australia

<sup>3</sup> Helmholtz Zentrum Dresden-Rossendorf, Helmholtz Institute Freiberg for Resources Technology, Freiberg, Germany

## 1 Introduction

Understanding the particularities of geosystems and geoscience data is critical for obtaining accurate and physically consistent inferences and predictions (Reichstein et al. 2019). Due to technological advances in capturing geoscience data, the archives of input data are large and ever growing (Sellars 2018). Geoscience data are obtained from a variety of sources, including remote sensing (e.g. hyperspectral satellite images, airborne geophysical surveys and high-quality aerial photography via drones), in situ sensors in close proximity to the phenomenon under investigation (e.g. cameras on conveyor belts, multiple sensors in the flotation cells, and soil pH sensors), direct observations and sampling during field campaigns (e.g. soil geochemical samples and drilling data), historical records and simulation data generated from process-based models. Inconsistency and poor quality are therefore often unavoidable; For instance, noise and missing values, measurement errors and analytical errors often accompany geochemical data (Grunsky 2010).

Geoscience processes and attributes vary significantly through time and space. Such heterogeneity is related to the spatial and temporal variation of soil types, rock types, land uses, vegetation types, climatic conditions and tectonic activities. The heterogeneity and non-stationarity of geosystems and geoscience data must be accounted for during modelling of geoscience variables across all points in space and time (Chilès and Delfiner 2012). Geoscience attributes are spatially and/or temporally auto- and cross-correlated (Goovaerts, 1997; Webster and Oliver, 2007) or show even more complex statistical and spatial patterns (Mariethoz and Caers 2015); For instance, a geochemical sample that shows a low proportion of magnesium oxide (MgO) is generally surrounded by locations that have similar MgO proportions. This sample and surrounding locations potentially share similar geological characteristics, such as bedrock geology or surficial quaternary units.

Extracting information from high-dimensional and large datasets far exceeds any human's abilities. Machine learning (ML) approaches are used to extract the hidden information from such datasets. However, ML algorithms may not be optimal when applied to geoscience data (Karpatne et al. 2019). The particularities of geosystems and geoscience data (e.g. big data, multi-source, multi-scale, high-dimensionality, poor quality data, limited sample size, paucity of ground-truth information, physics-based systems, importance of extreme cases, spatial and temporal heterogeneity, auto- and cross-correlations and complex uncertainty model) should be accounted for in the development of ML algorithms suitable for geoscience data. The most advanced ML algorithms are accurate when good-quality training data are abundant, but are seldom transparent, credible or interpretable. In particular, a lack of physical realism is problematic, as it makes ML algorithms potentially inaccurate in terms of extrapolations and less interpretable in terms of input parameters and predictor ranking (Reichstein et al. 2019). The majority of ML algorithms are based on the assumption of identically, independently distributed data. Such algorithms are not credible when applied to geoscience data (Schaeben et al. 2019). Fit-for-purpose ML algorithms should be able to capture dynamic (varying through time and space) multivariate spatial and/or temporal patterns of different scales and types.

Either current statistical learning algorithms can be amended to be consistent with the nature of geoscience data or new algorithms need to be developed; For instance, earth science data can be clustered to split the domain of study to account for the radically different behaviours of the natural phenomenon over the domain (useful for earth process discovery) and to simplify the subsequent modelling steps. However, to achieve this goal, a consistent clustering algorithm should be implemented. Non-spatial clustering techniques generally cluster observations based on their relationships in the feature space, so they do not have the means to consider auto- and cross-correlations of the regionalised variables. As a result of the lack of spatial awareness, the spatial coherence of the resulting clusters is not ensured (Fouedjio 2016a).

Incorporating coordinates as additional dimensions into the feature space and applying classical non-spatial clustering algorithms subsequently may lead to unsatisfactory results; For instance, two distal (in geographical space) points may belong to the same geological unit of interest or teleconnections in climate studies (Kawale et al. 2013). Spatial contiguity can also be enforced during the clustering process by imposing a proximity condition based on a graph organising the observations in the geographical space (Romary et al. 2015). Secchi et al. (2013) proposed a technique to cluster spatially dependent functional data using random Voronoi tessellations. In their proposed approach, the original data are replaced by some local representatives, and these local data are clustered subsequently. In addition, they achieved a model of uncertainty by implementing a bagging process. Another possibility is to apply the non-spatial clustering algorithms on a modified version of the dissimilarity matrix. Dissimilarities between observations are modified to take into account the spatial dependence (Oliver and Webster 1989; Bourgault et al. 1992; Fouedjio 2016a, b).

However, the aforementioned possibilities for improving the spatial awareness of a clustering algorithm are not suitable for recognising complex spatial patterns, objects and structures of different scales (which are not easily captured by two-point geostatistics) or their spatial distributions across the domain of study. Deep clustering technique such as those based on convolutional auto-encoders (Guo et al. 2017; Min et al. 2018) need a pre-processing step for multi-sensor imagery and vector data (e.g. interpolation and upscaling), and usually there is no sensitivity analysis on the effects of such pre-processing on the final results. The black-box nature of deep learning techniques makes the inference process somewhat difficult (Kuwajima et al. 2019). More importantly, the clustering of spatial mixed (e.g. continuous and categorical) and constrained (e.g. compositional) data has not been discussed.

The main objectives of this paper are to highlight the nature and characteristics of geoscience data and geosystems, the limitations of ML algorithms when it comes to spatially dependent data, and finally illustrate how classical ML algorithms can be amended to account for the particularities of spatial data through a case study aiming to improve the spatial awareness of spectral clustering.

Spectral clustering is a relatively recent clustering algorithm that relies on graph-theoretical concepts (Ng et al. 2001; von Luxburg 2007) but generally does not account for spatial relationships between observations. A new approach is introduced to generate a dissimilarity matrix computed from the distances between multivariate data events (e.g. geophysical, geochemical and geological spatial patterns) of different size and scales. Subsequently, spectral clustering will be used to cluster the input data based

on the novel dissimilarity matrix. Incorporating existing domain knowledge (e.g. prior knowledge on the size and geometry of geophysical, geochemical and geological spatial patterns, and reliability and data abundance for any sources of information) into models of this type generates more physically realistic outputs.

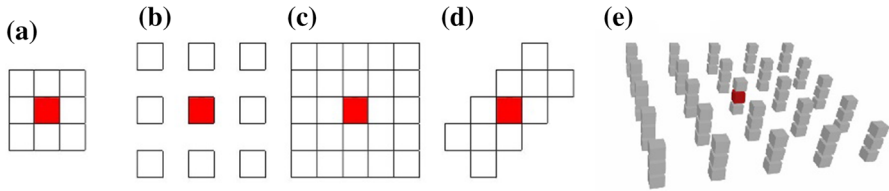
The proposed technique will be illustrated through one synthetic and one mining case study where the primary geometallurgical attributes are spatially clustered to split the deposit into different domains. These domains will simplify the subsequent metallurgical sampling and modelling steps.

Section 2 presents the proposed methodology for physically realistic and spatially aware spectral clustering. Section 3 illustrates the implementation of the new technique using synthetic and real cases. Finally, some conclusions and final thoughts are presented in Sect. 4.

## 2 Methodology

Spectral clustering algorithms cluster the eigenvectors derived from a similarity matrix of input data. Normally, a kernel function, e.g. a radial basis function (RBF) kernel, is used to generate the  $n \times n$  similarity matrix (where  $n$  is the number of input data). The kernel function ignores spatial information such as heterogeneity, auto- and cross-correlations, as well as complex spatial objects and patterns; For instance, in the case of the RBF, the kernel function is based on the pairwise squared Euclidean distance between observations in the feature space. Subsequently, a normalised Laplacian is defined from the non-spatial similarity matrix. After defining the eigensystem of the Laplacian, a classical clustering algorithm such as  $k$ -means is applied on the top  $m$  eigenvectors to obtain  $m$  clusters. To improve the spatial awareness of the spectral clustering, the following algorithm is proposed:

A set of  $n$  regionalised multivariate data  $Y = \{y(u_i) = [y^1(u_i), y^2(u_i), \dots, y^K(u_i)]; i = 1, \dots, n\}$  needs to be clustered into  $m$  subsets. The variables  $y^1(\cdot), y^2(\cdot), \dots, y^K(\cdot)$  can be a mixture of categorical and continuous spatial variables, and some of the continuous variables may be compositional. For ease of discussion, it is assumed that the first  $L, L \geq 0$ , variables form a composition. For each location  $u_i$ , the data event for the  $k$ th variable is defined as  $d_{E_k}(u_i) = \{y^k(u_{ij}); j = 1, \dots, E_k\}$ , which consists of the values of the  $k$ th variable,  $k = L + 1, \dots, K$ , at all the  $E_k$  nodes (including the node  $u_i$ ) in the neighbourhood. If a subset of the continuous variables is compositional,  $y^1(\cdot), \dots, y^L(\cdot)$ , then the data event for these variables is multivariate and defined as  $d_{E_Z}(u_i) = \{[y^1(u_{ij}), \dots, y^L(u_{ij})]; j = 1, \dots, E_Z\}$ . The overall data event at  $u_i$  is then given by  $d_E(u_i) = [d_{E_Z}(u_i), d_{E_{L+1}}(u_i), \dots, d_{E_K}(u_i)]$ . The parameters  $E_Z$  and  $E_k, k = L + 1, \dots, K$  control the order of spatial statistics (the size of the patterns) relevant for the composition and for the  $k$ th variable. Selecting  $E_k = E_Z = 1$  changes the proposed spatial algorithm to the classical non-spatial clustering. The order of spatial statistics and the geometry of the spatial pattern can be different for each source of information (e.g. mineralogy and geochemical compositions, rock type, alteration code, porosity, permeability and resistivity). Figure 1 shows examples of data event geometry useful for measuring pattern similarity at small scale, large



**Fig. 1** Examples of data event geometry useful for measuring pattern similarity at **a** small scale ( $E = 9$ ), **b** large scale ( $E = 9$ ), **c** a combination of all scales ( $E = 25$ ), **d** oriented ( $E = 11$ ), and **e** three-dimensional ( $E = 75$ )

scale, a combination of all scales, oriented, and three-dimensional, respectively. It is the responsibility of the expert users to define the most relevant geometry for the spatial patterns based on their prior knowledge.

To define a distance between two multivariate data events, the nature of the variables must be considered. If a subset  $Z$  composed of the first  $L$  variables is compositional, then the distance is given by

$$d_{\text{composition}}^Z \{d_{E_Z}(u_i), d_{E_Z}(u_j)\} = \frac{1}{E_Z d_{\text{max}}^Z} \sum_{e=1}^{E_Z} \sqrt{\frac{1}{2L} \sum_{r=1}^L \sum_{q=1}^L \left( \ln \frac{y^r(u_{ie})}{y^q(u_{ie})} - \ln \frac{y^r(u_{je})}{y^q(u_{je})} \right)^2}, \tag{1}$$

where  $d_{\text{max}}^Z$  is the compositional range, i.e. the longest Aitchison distance between compositional observations (Aitchison 1982), for the index set  $Z$ . If the  $k$ th variable,  $k \geq L + 1$ , is categorical, then the distance is defined as

$$d_{\text{categorical}}^k \{d_{E_k}(u_i), d_{E_k}(u_j)\} = \frac{1}{E_k} \sum_{e=1}^{E_k} \left( 1 - \delta(y^k(u_{ie}), y^k(u_{je})) \right), \tag{2}$$

where  $\delta(x, y) = 1$  if  $x = y$  and 0 otherwise. If the  $k$ th variable is continuous, then the distance is defined as

$$d_{\text{continuous}}^k \{d_{E_k}(u_i), d_{E_k}(u_j)\} = \frac{1}{E_k d_{\text{max}}^k} \sum_{e=1}^{E_k} \sqrt{[y^k(u_{ie}) - y^k(u_{je})]^2}, \tag{3}$$

where the normalising factor  $d_{\text{max}}^k$  is the range of the  $k$ th variable.

The total distance between two spatial mixed data events is the convex combination of the individual distances

$$d_{ij} \{d_E(u_i), d_E(u_j)\} = \mu_Z d_{\text{composition}}^Z + \sum_{k=L+1}^K \mu_k d^k, \tag{4}$$

where  $\mu_Z + \sum_{k=L+1}^K \mu_k = 1; \mu_Z, \mu_k \geq 0$ , and

$$d^k = \begin{cases} d_{\text{categorical}}^k & \text{if the } k\text{th variable is categorical} \\ d_{\text{continuous}}^k & \text{if the } k\text{th variable is continuous} \end{cases}$$

By construction,  $d_{ij} \in [0, 1]$ . The function  $d_{ij}$  satisfies the following conditions:  $d_{ij} \geq 0$ ,  $d_{ij} = 0$  if and only if  $i = j$ ,  $d_{ij} = d_{ji}$ , and  $d_{ij} \leq d_{ik} + d_{kj}$ . The weights  $\mu_Z$  and  $\mu_k$  account for the multiple-point spatial dependence between mixed variables. It is highly recommended that the geometries of the spatial patterns and dependency weights be defined by geoscientists based on their expert domain knowledge.

The radial basis function is used to define a symmetric similarity matrix  $A_{n \times n}$  from the spatially aware distance matrix as

$$A_{ij} = (1 - \delta(i, j)) \exp\left(\frac{-d_{ij}^2}{2\sigma^2}\right). \quad (5)$$

The scaling parameter  $\sigma$  controls the deceleration of the similarity  $A_{ij}$  with distance  $d_{ij}$ , and several approaches have been introduced for choosing it automatically (Ng et al. 2001; Zelnik-Manor and Perona 2004). In this study, to select a practical value for  $\sigma$ , a distribution of the  $n$  closest distances is generated from the distance matrix (minimum for each row, excluding diagonal values), and the 99th percentile is selected as  $\sigma$ . The symmetric normalised Laplacian is subsequently defined as

$$L = D^{-1/2} A D^{-1/2}, \quad (6)$$

where  $D$  is the diagonal matrix whose  $i$ th element is the sum of the entries in the  $i$ th row of  $A$ . The Laplacian  $L$  is symmetric, and its eigenvectors can be chosen to be pairwise orthogonal. In spectral clustering implementation, the eigengap heuristic is often used to find the number of clusters  $m$ . The  $m$  largest eigenvalues represent sub-regions of the study area that share similar multivariate spatial patterns. The eigengap  $\delta_m = |\lambda_{m+1} - \lambda_m|$  is the absolute difference between the  $(m+1)$ th and  $m$ th largest eigenvalues of  $L$ . The eigenvectors  $x_1, \dots, x_m$  (counted according to their multiplicity) corresponding to the  $m$  largest eigenvalues are used to form the  $n \times m$  matrix  $X = [x_1 \dots x_m]$ . Based on matrix perturbation theory, the subspace spanned by  $X$  is stable if  $\delta_m$  is sufficiently large. The rows of  $X$  are normalised to unit Euclidean length. The resulting points on the  $m$ -dimensional unit sphere can then be clustered via a distortion minimisation technique such as  $K$ -means clustering. Finally, the original observation  $\mathbf{y}(u_i)$  is assigned to a cluster  $s \in (1, 2, \dots, m)$  if and only if the  $i$ th point was assigned to cluster  $s$ . Theoretical considerations on spectral clustering can be found in Ng et al. (2001) and von Luxburg (2007). There is no assumption on the underlying model for the proposed clustering algorithm, thus a large variety of different indices (Charrad et al. 2014) can be used to pick the number of clusters and optimise the input parameters including the geometry of data events,  $E_Z$ ,  $E_k$ ,  $\mu_Z$ , and  $\mu_k$ ; For instance, the silhouette metric (ranging from  $-1$  to  $+1$ ) that measures the similarity of an object to its own cluster and dissimilarity from the other clusters can be used to tune the input parameters. The input parameters for the proposed technique can also be selected by experts in a knowledge-driven way. In this study, this approach

is implemented. The input parameters enable experts to use their existing domain knowledge. The resultant clusters are highly dependent on the experts' knowledge and decisions. Each geological structure determines the choice of geometry for the corresponding spatial pattern in a multivariate data event; For instance, a geological structure such as a weathering profile may be captured using a spatial pattern with a greater vertical than horizontal extent. Similar decisions need to be made for alteration zones and intensities, faults and folds, and geochronological order. Reliability and data abundance for any sources of information can also be handled via different weighting configurations. Such flexibility in terms of the physical realism of the input parameters leads to an interpretable and reliable unsupervised model.

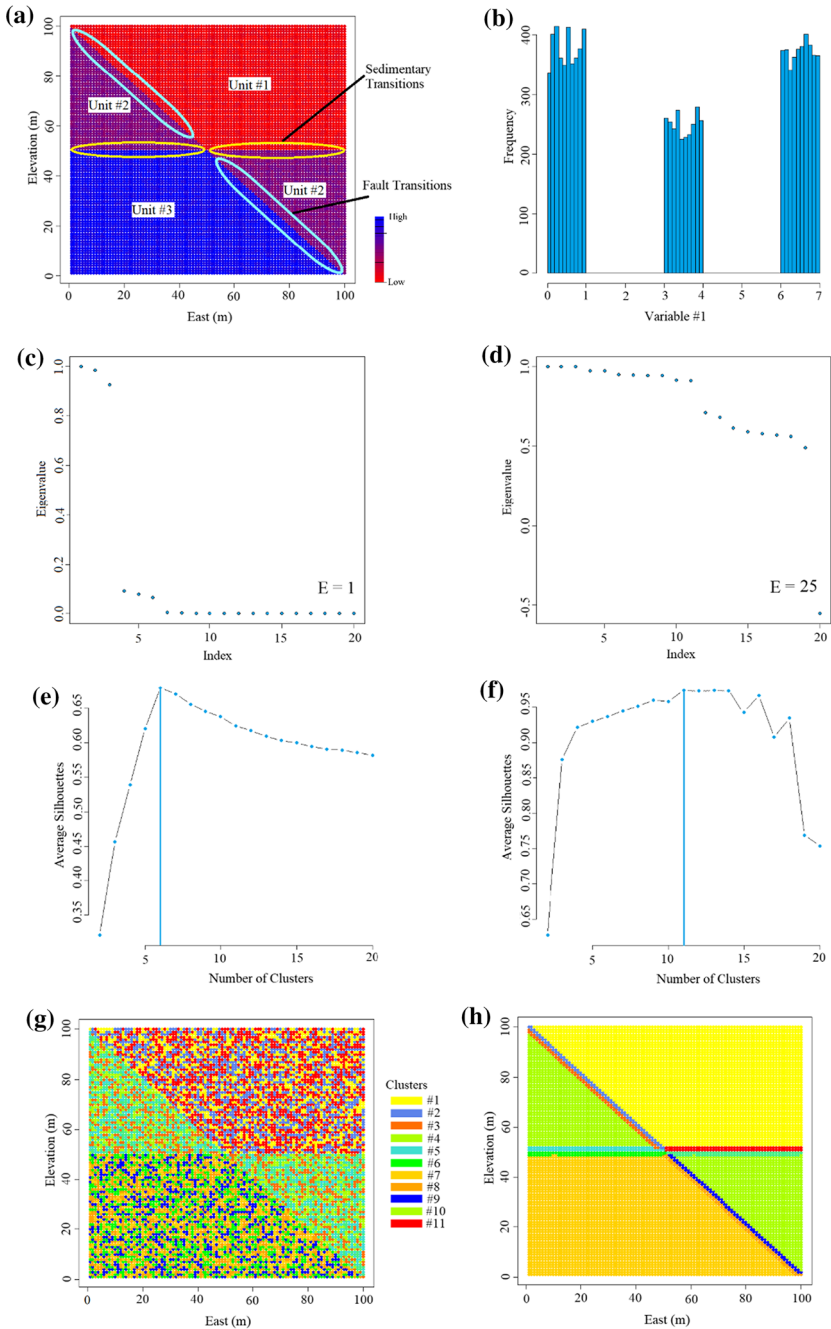
It should be noted that some pre-processing might be required. If the input data do not lie on a regular grid, they should first be migrated to the closest nodes of a regular grid of suitable resolution (some inputs such as satellite images are already regularly spaced) or be rasterised using geostatistical simulation techniques. Similarly, missing values and empty cells should be imputed first, and multiple-point simulations are recommended for this purpose. Moreover, dissimilarity measures for the data located at the margins of the study area are based on a lower order of spatial statistics (part of the data event is located outside the known region). For the cases where full pattern matching is needed, the margins can be dropped from calculations.

### 3 Experiments

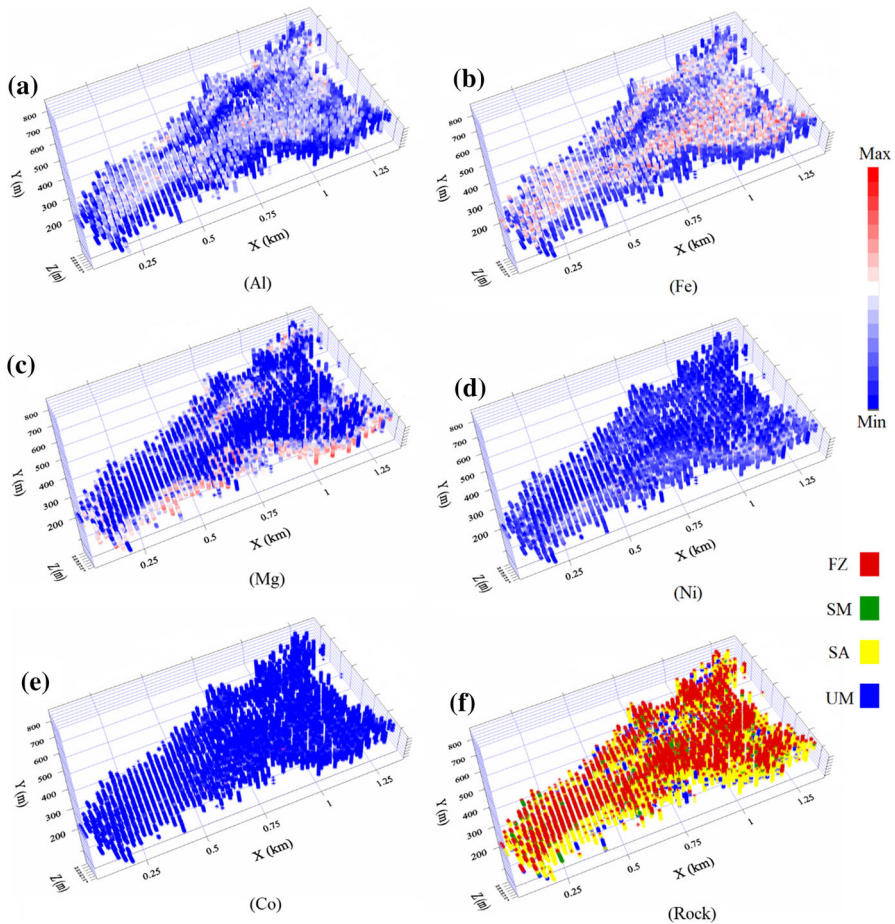
#### 3.1 Synthetic Case Study

The performance of the proposed method was tested using a synthetic case study (Fig. 2). Figure 2a shows a geological cross-section. Three geological units, two types of sedimentary transition and two types of fault transition (rock type #1 to #2 and rock type #2 to #3) are indicated in this figure. Three samples from a uniform distribution (with different centres) were combined to define the input variable (Fig. 2b). Classical spectral clustering ( $E_k = 1$ , model M1) was implemented to assess the gain in the implementation of a spatially and physically aware learner. The geometry in Fig. 1c ( $E_k = 25$ ) was selected to build the spatially aware distance matrix (model M2). Figure 2c and d show the large eigenvalues of the symmetric normalised Laplacians for models M1 and M2, respectively. The eigengap heuristic suggests three clusters (jump after the third largest eigenvalue) for the M1 and 11 for the M2 model. Comparing Fig. 2e and f reveals the superior performance (larger silhouette value) of the proposed algorithm in terms of generating tight clusters. There is a consistency between the eigengap heuristic and the silhouette metric in this case; For instance, both of these methods suggest 11 clusters for the M2 model. The final clustered maps are shown in Fig. 2g, h. The proposed method recognised the three geological units and the four types of transitions. In addition, M2 recognised the different sides of transitions, which is useful for some geological phenomenon such as uranium roll-front or skarn





**Fig. 2** **a** Cross-section of the input data, **b** histogram of the input data, large eigenvalues of the Laplacian matrix for model **c** M1 and **d** M2, optimum number of clusters (based on silhouette metric) for model **e** M1 and **f** M2, and final clustered map for model **g** M1 and **h** M2

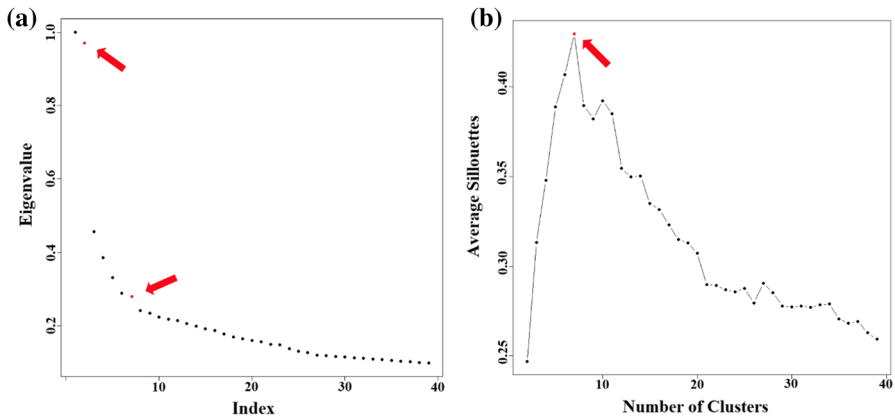


**Fig. 3** Three-dimensional representation of the input data, **a–e** continuous variables of interest, and **g** categorical variable (rock type)

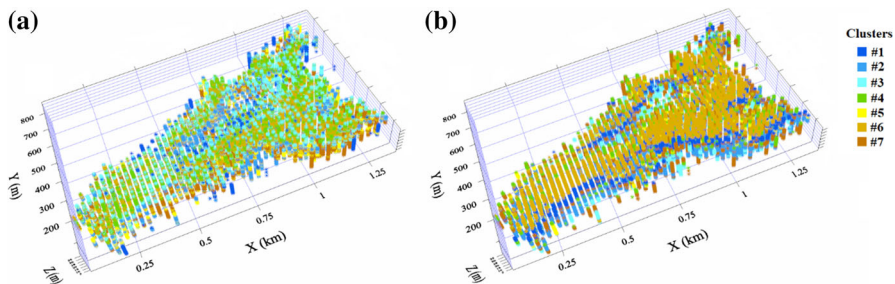
mineralisation. The proposed method is also robust in terms of removing unstructured noise and improving the spatial coherency of the final clusters (Fig. 2g, h).

### 3.2 Mining Data

Murrin Murrin East (MME) is a nickel–cobalt laterite deposit located in Western Australia. At MME, nickel laterite deposits occur as laterally extensive, undulating blankets of mineralisation with strong vertical trends covering basement ultramafic rocks (Murphy 2003). In total, 17,377 samples (of length 1 m) from 920 regularly spaced RC holes (25 m × 25 m) make up the database for this study (Fig. 3). Variables



**Fig. 4** Selecting the number of clusters using the **a** eigengap heuristic and **b** average silhouette



**Fig. 5** Spatial distribution of the clustered data: **a** non-spatial clustering ( $E_k = E_Z = 1$ ) and **b** spatial clustering ( $E_k = E_Z = 75$ )

of interest consist of one categorical variable, viz. rock type: ferruginous (FZ), smectite (SM), saprolite (SM) and ultramafic (UM), and geochemical compositions: three major (Fe, Al and Mg) and two target (Ni and Co) elements, plus Rest to achieve closure in the compositional data.

The geometry of the three-dimensional data event in Fig. 1e was selected to measure the distances. In this study, the same parameters were used for geological information and compositional geochemistry ( $E_k = E_Z = 75$ ). Weights  $\mu_Z = 0.7$  and  $\mu_7 = 0.3$  were defined to give 70% importance to compositional information and 30% to rock types. Equation 1 was used to measure the distance between spatial patterns of geochemical compositions, and Eq. (2) to measure the distance between spatial patterns of lithological units. Classical spectral clustering ( $E_k = E_Z = 1$ ,  $\mu_Z = 0.7$  and  $\mu_7 = 0.3$ ) was also implemented to assess the gain in the implementation of the spatial spectral clustering for geometallurgical domaining.

The eigengap heuristic suggested two (with a major jump after the second largest eigenvalue) while the average silhouette index showed seven clusters as the best option (Fig. 4a, b). The first two eigenvalues are related to the dominant geological units, FZ and SA (Fig. 3f). To provide more detail, boreholes were clustered into seven spatially homogeneous regions. Figure 5a and b show the spatial distribution of the final seven clusters for the non-spatial and spatial spectral clustering, respectively. The non-spatial clustering generated scattered domains (less spatial continuity), while the spatial clustering recognised domains with complex structure; For instance, cluster #1 shows complex curvilinear structure that is not easy to capture by two-point geostatistics.

Compositional characteristics such as closed geometric means, total compositional variances and rock proportions (Pawlowsky-Glahn et al. 2015) of the generated spatial clusters are presented in Table 1. Cluster #6 is rich in Fe, depleted in Mg and mainly composed of ferruginous rocks. Cluster #5 shows high levels of Co and Ni mineralisation and mainly consists of smectite units. Fresh ultramafic rocks are mainly located in cluster #3. This cluster shows the lowest level of global compositional dispersion, which might be related to the fact that weathering has not reached to this depth. The spatial distribution and compositional characteristics of the domains generated by the spatial spectral clustering are consistent with the current geological understanding of this deposit (Talebi et al. 2017, 2019).

## 4 Conclusions

Traditional statistical learning techniques do not fully account for the particularities of geosystems and geoscience data. These techniques must thus be amended to account for the characteristics of geoscience data, or new spatial learners must be developed. A novel methodology is proposed herein to improve the physical realism and spatial awareness of spectral clustering. This spatial spectral clustering approach allows the use of existing domain knowledge to select the input parameters (e.g. geometry of the patterns, order of spatial statistics, and weights for the convex combination of distance matrices). Clusters generated via the proposed algorithm are homogeneous (similar multivariate spatial patterns) up to a selected order of spatial statistics, thus presenting a richer view of stationarity than variogram-based or non-spatial clustering techniques. Many attributes of different natures (continuous, categorical and constrained) can be used as inputs to the proposed clustering technique. The results from the synthetic and mining datasets proved the usefulness of the technique. The proposed method will be developed further for clustering of multisensor imagery and vector data simultaneously.

**Table 1** Statistical characteristics of each spatially clustered region

	Al	Fe	Mg	Rest	Ni	Co	totvar	FZ	SM	SA	UM
Cluster #1	4.60	26.32	1.01	67.33	0.69	0.05	4.425	0.658	0.088	0.239	0.015
Cluster #2	0.94	11.19	10.84	76.21	0.79	0.03	1.466	0.005	0.004	0.982	0.009
Cluster #3	0.90	8.17	12.86	77.58	0.47	0.02	1.176	0.002	0.002	0.014	0.982
Cluster #4	5.57	28.89	0.68	64.28	0.54	0.04	3.844	0.890	0.037	0.070	0.003
Cluster #5	2.36	23.68	2.78	69.84	1.26	0.08	2.604	0.011	0.937	0.049	0.003
Cluster #6	8.00	29.79	0.35	61.61	0.23	0.02	1.919	0.957	0.033	0.008	0.002
Cluster #7	1.52	11.66	8.62	77.50	0.67	0.03	2.599	0.001	0.002	0.995	0.002

**Acknowledgements** This study was funded through CSIRO’s Deep Earth Imaging Future Science Platform.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aitchison J (1982) The statistical analysis of compositional data. *J R Stat Soc Ser B* 44:139–177
- Bourgault G, Marcotte D, Legendre P (1992) The multivariate (co)variogram as a spatial weighting function in classification methods. *Math Geol* 24:463–478
- Charrad M, Ghazzali N, Boiteau V, Niknafs A (2014) NbClust: an R package for determining the relevant number of clusters in a data set. *J Stat Softw* 61:1–36
- Chilès J-P, Delfiner P (2012) *Geostatistics: modeling spatial uncertainty*, 2nd edn. Wiley, New York
- Fouedjio F (2016a) A hierarchical clustering method for multivariate geostatistical data. *Spat Stat* 18:333–351. <https://doi.org/10.1016/j.spasta.2016.07.003>
- Fouedjio F (2016b) A clustering approach for discovering intrinsic clusters in multivariate geostatistical data. In: Perner P (ed) *Machine learning and data mining in pattern recognition*. Springer International, Cham, pp 491–500
- Goovaerts P (1997) *Geostatistics for natural resources evaluation*. Oxford University Press, New York
- Grunsky EC (2010) The interpretation of geochemical survey data. *Geochem Explor Environ Anal* 10:27–74
- Guo X, Liu X, Zhu E, Yin J (2017) Deep clustering with convolutional autoencoders. In: Liu D, Xie S, Li Y et al (eds) *Neural information processing*. Springer International, Cham, pp 373–382
- Karpatne A, Ebert-Uphoff I, Ravela S et al (2019) Machine learning for the geosciences: challenges and opportunities. *IEEE Trans Knowl Data Eng* 31:1544–1554
- Kawale J, Liess S, Kumar A et al (2013) A graph-based approach to find teleconnections in climate data. *Stat Anal Data Min* 6:158–179
- Kuwajima H, Tanaka M, Okutomi M (2019) Improving transparency of deep neural inference process. *Prog Artif Intell* 8:273–285
- Mariethoz G, Caers J (2015) *Multiple-point geostatistics: stochastic modeling with training images*. Wiley, New York
- Min E, Guo X, Liu Q et al (2018) A survey of clustering with deep learning: from the perspective of network architecture. *IEEE Access* 6:39501–39514
- Murphy M (2003) *Geostatistical optimisation of sampling and estimation in a nickel laterite deposit*. Edith Cowan University (unpublished)
- Ng AY, Jordan MI, Weiss Y (2001) On spectral clustering: analysis and an algorithm. In: *Proceedings of the 14th international conference on neural information processing systems: natural and synthetic*. MIT Press, Cambridge, pp 849–856
- Oliver MA, Webster R (1989) A geostatistical basis for spatial weighting in multivariate classification. *Math Geol* 21:15–35
- Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2015) *Modeling and analysis of compositional data*. Wiley, Caldwell
- Reichstein M, Camps-Valls G, Stevens B et al (2019) Deep learning and process understanding for data-driven Earth system science. *Nature* 566:195–204
- Romary T, Ors F, Rivoirard J, Deraisme J (2015) Unsupervised classification of multivariate geostatistical data: two algorithms. *Comput Geosci* 85:96–103
- Schaeben H, Kost S, Semmler G (2019) Popular raster-based methods of prospectivity modeling and their relationships. *Math Geosci*. <https://doi.org/10.1007/s11004-019-09808-6>

- Secchi P, Vantini S, Vitelli V (2013) Bagging Voronoi classifiers for clustering spatial functional data. *Int J Appl Earth Obs Geoinf* 22:53–64
- Sellars SL (2018) “Grand challenges” in big data and the earth sciences. *Bull Am Meteorol Soc* 99:9ES95–ES98
- Talebi H, Lo J, Mueller U (2017) A hybrid model for joint simulation of high-dimensional continuous and categorical variables. In: Gómez-Hernández JJ, Rodrigo-Ilarri J, Rodrigo-Clavero ME et al (eds) *Geostatistics Valencia 2016*. Springer International, Cham, pp 415–430
- Talebi H, Mueller U, Tolosana-Delgado R, van den Boogaart KG (2019) Geostatistical simulation of geochemical compositions in the presence of multiple geological units: application to mineral resource evaluation. *Math Geosci* 51:129–153
- von Luxburg U (2007) A tutorial on spectral clustering. *Stat Comput* 17:395–416
- Webster R, Oliver MA (2007) *Geostatistics for environmental scientists*, 2nd edn. Wiley, Hoboken
- Zelnik-Manor L, Perona P (2004) Self-tuning spectral clustering. In: *Proceedings of the 17th international conference on neural information processing systems*. MIT Press, Cambridge, pp 1601–1608