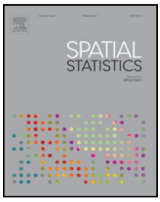
Spatial Statistics 35 (2020) 100407

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta



A robust hierarchical clustering for georeferenced data

Pierpaolo D'Urso *, Vincenzina Vitale

Dipartimento di Scienze Sociali ed Economiche, Sapienza Università di Roma, Italy

ARTICLE INFO

Article history:
Received 27 August 2019
Received in revised form 22 December 2019
Accepted 29 December 2019
Available online 9 January 2020

Keywords:
Agglomerative hierarchical clustering
Geostatistics
Kernel function
Robust dissimilarity measure
Multivariate spatial data
Top soil heavy metal concentrations

ABSTRACT

The detection of spatially contiguous clusters is a relevant task in geostatistics since near located observations might have similar features than distant ones. Spatially compact groups can also improve clustering results interpretation according to the different detected subregions. In this paper, we propose a robust metric approach to neutralize the effect of possible outliers, *i.e.* an exponential transformation of a dissimilarity measure between each pair of locations based on non-parametric kernel estimator of the direct and cross variograms (Fouedjio, 2016) and on a different bandwidth identification, suitable for agglomerative hierarchical clustering techniques applied to data indexed by geographical coordinates. Simulation results are very promising showing very good performances of our proposed metric with respect to the baseline ones. Finally, the new clustering approach is applied to two real-word data sets, both giving locations and top soil heavy metal concentrations.

1. Introduction

In the last decades, the growing amount of available spatial data imposes new research challenges such as the identification of spatially connected homogeneous subdomains for which data locations belonging to the same group show proximity both in the attribute and in the geographical space. The well known clustering techniques, applied to georeferenced data, fail to accomplish these two aspects. Therefore, in literature, to enforce the spatial connectivity of the resulting clusters, the main proposals consist of adapting the existing non spatial clustering techniques in a spatial context. As a result, a group of techniques takes into account spatial information using the geographical coordinates as additional variables with the disadvantage of producing too scattered clusters.

* Corresponding author.
E-mail address: pierpaolo.durso@uniroma1.it (P. D'Urso).

https://doi.org/10.1016/j.spasta.2020.100407
2211-6753/

Other approaches refer to the adjustment of the dissimilarity matrix of the data, adequately weighted by the variogram (Oliver and Webster, 1989) and to spatially constrained clustering algorithms (Romary et al., 2015). In this latest cited work, two algorithms are proposed that are model free and can handle large volumes of multivariate, irregularly spaced data. Both are adaptations of agglomerative hierarchical and spectral algorithms. The spatial coherence is ensured by a proximity condition imposed for two clusters to merge. A new approach has been recently proposed by Fouedjio (2016) based on an agglomerative hierarchical clustering: the dissimilarity measure between each pair of locations is computed by means of a non-parametric kernel estimator of the multivariate spatial dependence structure of data, *i.e.* a non-parametric kernel estimator of direct and cross variograms. The proposed clustering approach is model-free and could be adapted to irregularly spaced data, ensuring spatially contiguous clusters without imposing any geometrical constraints. We can refer to the last two cited papers for a very interesting review of existing approaches in literature, both model free and model based.

Another relevant issue, for clustering task and for data mining in general, concerns the outliers detection, that is to identify units that markedly deviate from the rest of data (Barnett and Lewis, 1994). García-Escudero et al. (2010) point out the existing relevant connection between cluster analysis and robust statistics and, in general, it is well known that robust cluster analysis is strictly related to the scope of the anomaly detection.

Moreover, as well pointed out by García-Escudero et al. (2008), "the precise detection of the outliers is an important task due to the serious troubles they introduce in standard clustering procedures as well as the appealing interest that outliers could have by themselves after explaining why they depart from general behavior". In this regard, we argue about the massive use of data transformation and smoothing techniques regardless of whether it is the target of the analysis. If the aim is to identify anomalies in a system, such as the first alert or alarm in the monitoring of high traffic levels, data transformation is exactly what it has to be avoided.

In the context of our clustering approach, belonging to the wider class of the heuristics\ exploratory techniques (Gordon, 1999; Everitt et al., 2011), typically model-free (therefore, no mixture-based), the outlier's definition is not related to a model of data generation. As pointed out by Barnett (1978), "if the purpose is non-model-specific location estimation, preliminary rejection of outlier following a test of discordancy is irrelevant. We need an estimator which accommodates outliers; that is robust against their presence. An example is the median, or some other appropriate trimmed mean". In the same work, among the four basic ways of handling outliers, Barnett specifies the "Accommodation" method, essentially based on robust statistics for which there need be no specification of an initial model.

Therefore, as far as model free clustering techniques are concerned, the outliers could be a group of observations (smaller with respect to the main clusters) that differ from the proper clusters or, alternatively, could be represented by isolated points, each forming its own group (García-Escudero et al., 2003). In this context, the outlier is identified according to its distance with respect to the bulk of data. For an accurate dissertation about the role and definition of the outliers in cluster analysis as well as their representation in terms of distance you could refer to the previous cited work.

In the spatial context, outliers do not necessarily deviate from the whole data set, it could also be viewed as a local anomaly whose non-spatial attribute values are extreme with respect to its neighbour (Cerioli and Riani, 1999; Shekhar and Chawla, 2003). As pointed out by Chen et al. (2008), traditional outliers detection techniques may not be directly used to identify abnormal spatial patterns due to the particular features of spatial data: they have more complex structures and, above all, they are often characterized by spatial contiguity and autocorrelation of the nearest locations.

Taking into account spatial relationships between data locations, then the outliers identification focuses on the non-spatial attributes values. In literature, spatial outliers detection methods are mainly devoted to detect single attribute outliers. The graphic approaches use variogram clouds and pocket plots (Haslett et al., 1991; Pannatier, 2012), the quantitative approaches are based on tests for detecting anomalies; the main representative techniques are the scatterplot and the Moran scatterplot (Anselin, 1995). In Cerioli and Riani (1999), to identify spatial outliers, a forward search algorithm, ordering the observations from those most in agreement with a specified autocorrelation model to those least in agreement with it, is proposed.

Among clustering algorithms, the density-based algorithm DBSCAN (Density-Based Spatial Clustering of Applications with Noise), proposed by Ester et al. (1996), explicitly separates clusters of

high density from clusters of low density, marking as outliers points that lie alone in low-density regions. It is able to find arbitrarily shaped clusters, without specifying the number of clusters in the data a priori. The main disadvantages are that it does not work well on high-dimensional data and it is more sensitive to the choice of its parameters that has to be defined by users: the minimum number of data points needed to determine a single cluster and how close points should be to each other to be considered a part of a cluster.

For other interesting references on quantitative techniques and neighbourhood-based approaches see Liu et al. (2001), Shekhar et al. (2001), Lu et al. (2003) and Chen et al. (2008).

In this paper, with the aim of handling spatial outliers in clustering procedure, we propose a new modified version of the kernel estimator proposed by Fouedjio (2016) suitable for spatial data in presence of outliers.

In particular, following a non-parametric approach, we propose a robust version of the dissimilarity index suggested by Fouedjio (2016) for spatial data. Properly, the robust dissimilarity index is based on an exponential transformation of the cited index and on a new criterion for the bandwidth selection. It can be used fruitfully in the agglomerative hierarchical clustering so that the clustering methods based on the proposed robust dissimilarity measure are capable to neutralize the effect of the outliers in the clustering process and, therefore, they are able to respect the natural structure of spatial data.

The outline of the article is as follows. In Section 2, the robust non parametric kernel estimator is introduced together with a little description of the agglomerative hierarchical clustering techniques; in Section 3 simulation results are shown in detail while in Section 4 two applications to real data are proposed. In Section 5 we address some conclusions and open research problems.

2. Methodology

As already pointed out, the aim of this work is to provide a modified version of the dissimilarity measure proposed by Fouedjio (2016) in order to make it suitable in presence of spatial outliers. In the cited work, the dissimilarity measure between each pair of locations is computed by means of a non-parametric kernel estimator of the multivariate spatial dependence structure of data, *i.e.* a non-parametric kernel estimator of direct and cross variograms. Properly, the next paragraphs deeply focus on the theoretical aspects of the robust approach based on the above non parametric dissimilarity measure.

2.1. Dissimilarity measure

Let $G \subset \Re^d$, with $d \geq 1$, a fixed continuous spatial domain and let $\{Z_1 \dots Z_p\}$ p standardized variables of interest measured at a set of distinct locations $\{\mathbf{s}_t \in G\}_{t=1}^n$.

The non-parametric kernel estimator of the direct and cross variograms, at two locations $\mathbf{x} \in G$ and $\mathbf{y} \in G$, proposed by Fouedjio (2016), has the following mathematical form:

$$\widehat{\gamma}_{ij}(\mathbf{x}, \mathbf{y}; \lambda) = \frac{\sum_{k,l=1}^{n} K_\lambda((\mathbf{x}, \mathbf{y}), (\mathbf{s}_k, \mathbf{s}_l))(Z_i(\mathbf{s}_k) - Z_i(\mathbf{s}_l))(Z_j(\mathbf{s}_k) - Z_j(\mathbf{s}_l))}{2\sum_{k,l=1}^{n} K_\lambda((\mathbf{x}, \mathbf{y}), (\mathbf{s}_k, \mathbf{s}_l))} \mathbb{1}_{\{\mathbf{x} \neq \mathbf{y}\}} \tag{1}$$

where $(i, j) \in \{1, \dots, p\}^2$.

The non negative kernel function $K_\lambda((\mathbf{x}, \mathbf{y}), (\mathbf{s}_k, \mathbf{s}_l))$, with constant bandwidth parameter $\lambda > 0$, is equal to $K_\lambda(\|\mathbf{x} - \mathbf{s}_k\|)K_\lambda(\|\mathbf{y} - \mathbf{s}_l\|)$.

Thus, the normalized dissimilarity between two sample location, $\mathbf{s}_k$ and $\mathbf{s}_l$, is defined as:

$$d_\lambda(\mathbf{s}_k, \mathbf{s}_l) = \frac{\sum_{i,j=1}^{p} |\widehat{\gamma}_{ij}(\mathbf{s}_k, \mathbf{s}_l; \lambda)|}{max_{(k,l)\in\{1,\dots,n\}^2} \sum_{i,j=1}^{p} |\widehat{\gamma}_{ij}(\mathbf{s}_k, \mathbf{s}_l; \lambda)|}. \tag{2}$$

The kernel function in Eq. (1) weighs data locations in order to assign more weight to all data locations closer to the target one.

Among the most known kernel functions, the Epanechnikov kernel (Wand and Jones, 1995) has been chosen because it guarantees compact support thus reducing computational burden. The Epanechnikov kernel function is equal to $K_\lambda(\|\mathbf{x} - \mathbf{s}\|) = \frac{3}{4\lambda}(\lambda^2 - \|\mathbf{x} - \mathbf{s}\|^2) \cdot \mathbb{1}_{\|\mathbf{x}-\mathbf{s}\| \leq \lambda}$.

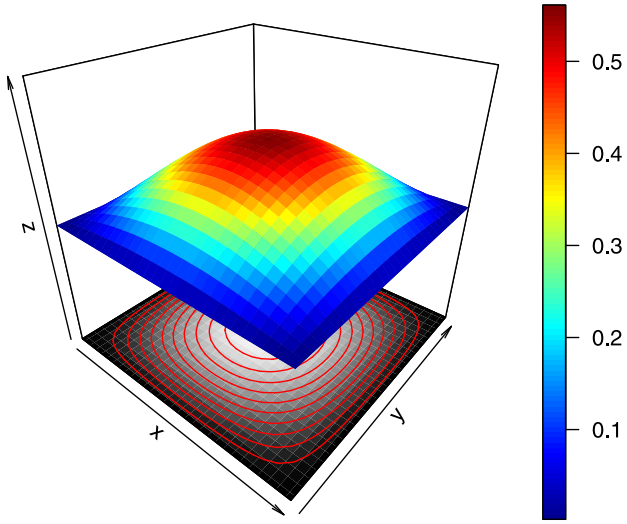


Fig. 1. Bivariate Epanechnikov kernel density.

An example of bivariate density is given in Fig. 1. It is worth noting that the selection of the bandwidth parameter λ is of greater interest: too small values do not include enough data locations in the kernel support leading to spatially non-contiguous clusters. On the contrary, too large values over-smooth the spatial dependence structure leading to clusters that may not reflect the true underlying structure. The next paragraph focuses on its appropriate choice.

2.2. Kernel bandwidth

Following Fouedjio (2016), λ is chosen such that the kernel support contains at least 35 observations. Therefore, it corresponds to the maximum distance from the 35th neighbour location. In this work, let $\mathbf{Z_x} = (Z_1, \ldots, Z_k, \mathbf{x})$, $\mathbf{Z_y} = (Z_1, \ldots, Z_k, \mathbf{y})$, $\mathbf{Z_{s_k}} = (Z_1, \ldots, Z_k, \mathbf{s_k})$ and $\mathbf{Z_{s_l}} = (Z_1, \ldots, Z_k, \mathbf{s_l})$ be the vectors of $K + 2$ dimensions including coordinates as additional variables, the distance used for bandwidth identification is based on the euclidean norm between the above vectors. As a consequence, the kernel function is modified as follows:

$$K_\lambda((\mathbf{Z_x}, \mathbf{Z_y}), (\mathbf{Z_{s_k}}, \mathbf{Z_{s_l}})) = K_\lambda(\|\mathbf{Z_x} - \mathbf{Z_{s_k}}\|) K_\lambda(\|\mathbf{Z_y} - \mathbf{Z_{s_l}}\|).$$

In other words, here, neighbour's definition is adapted to the scope of outliers detection, balancing the needing of ensuring geographical proximity and that of detecting anomalies in the corresponding attribute space.

2.3. Robust approach

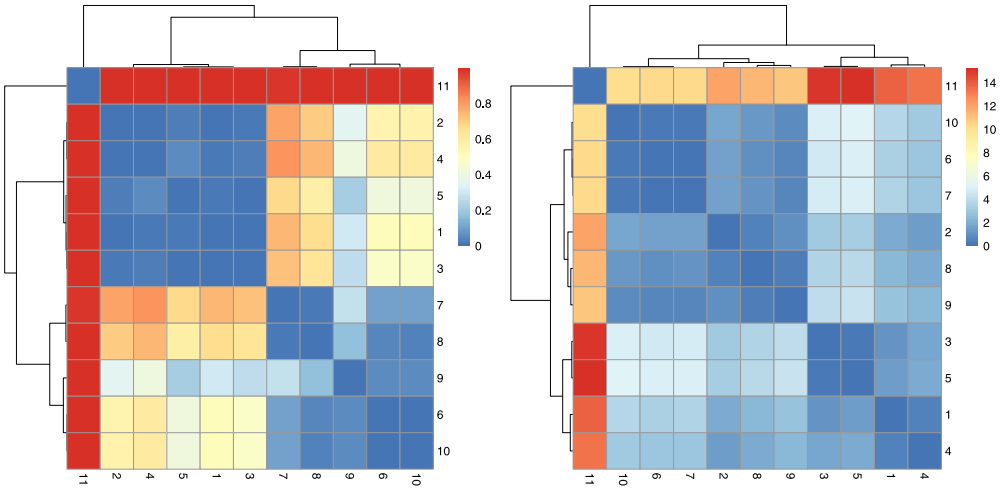
To achieve robustness in presence of spatial outliers, we propose the following exponential transformation of the dissimilarity index of Eq. (2) (Wu and Yang, 2002; D'Urso and De Giovanni, 2014):

$$d_{exp;\lambda}(\mathbf{s}_k, \mathbf{s}_l) = 1 - exp\{-\beta \cdot d^2_\lambda(\mathbf{s}_k, \mathbf{s}_l)\} \tag{3}$$

where we choose β as:

$$\beta = \left(Median\left(|d^2_\lambda(\mathbf{s}_k, \mathbf{s}_l) - M|\right)\right)^{-1} \tag{4}$$

for $k, l = 1, \ldots, n$ and $k < l$ (or $k > l$) and M the median of the squared dissimilarity indexes.



(a) Robust dissimilarities (b) Non-robust dissimilarities

Fig. 2. The effect of the exponential transformation on dissimilarities between units.

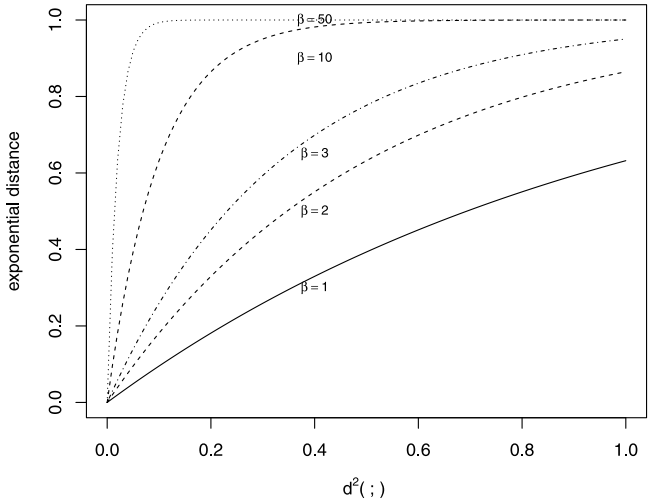


Fig. 3. Effect of the parameter β on the smoothed dissimilarity index (Eq. (3)).

The role of the exponential transformation (Eq. (3)), that lies between [0, 1], could be better understood if we take into account a simple example. Suppose we know that data structure is composed of two groups of five units each (units 1–5 belong to the former, units 6–10 to the latter) and suppose there is one outlier (the eleventh unit); the corresponding two maps of dissimilarities in Fig. 2 show that the robust metric (Fig. 2(a)) distinguishes two groups and a third one composed of the outlier (the unit 11); the non-robust metric (Fig. 2(b)), sensible to the presence of the outlier, creates a big group of the ten units and another one composed of the outlier.

As far as β parameter is concerned, it is usually chosen as the inverse of a measure of variability of the data and its effect on the exponential transformation (Eq. (3)) is illustrated in Fig. 3. In the presence of low variability of the data, increasing dissimilarities receive a weight lower than in the

case of high variability. For further insights on the definition and role of β, see Wu and Yang (2002) and D'Urso and De Giovanni (2014).

The resulting robust dissimilarity matrix will be used later in the agglomerative hierarchical clustering algorithms to handle spatial outliers, as shown in detail in the next sections.

2.4. The agglomerative hierarchical clustering

Given a dissimilarity matrix, an agglomerative hierarchical clustering algorithm (Everitt et al., 2011) could be applied in order to identify homogeneous groups of data locations. The agglomerative approach is a "bottom up" approach: each sample location starts in its own cluster; at each step of the hierarchy, the two clusters that are the most similar are combined into a new bigger cluster until all sample locations are merged in just one big cluster. The result is a tree-based representation of the objects, named dendrogram. At each step, the distances (dissimilarities) between the new formed cluster and each of the old clusters can be computed in different ways; the three most known linkage criteria, also used in this work, are named: *single linkage, average linkage and complete linkage*.

In the single linkage hierarchical clustering, the distance between two clusters G and H equals the minimum distance (the lowest dissimilarity) between any two members of the two clusters:

$$d_{\text{single}}(G, H) = \min_{i \in G,\, j \in H} d_{ij}.$$

In the average linkage hierarchical clustering, the distance between clusters G and H equals the arithmetic mean distance (the average dissimilarity) between all members in the two cluster members:

$$d_{\text{average}}(G, H) = \frac{1}{n_G \cdot n_H} \sum_{i \in G,\, j \in H} d_{ij}.$$

In the complete linkage hierarchical clustering, the distance between clusters G and H equals the maximum distance (the highest dissimilarity) between any two members of the two clusters:

$$d_{\text{complete}}(G, H) = \max_{i \in G,\, j \in H} d_{ij}.$$

It is well known that single linkage criterion tends to produce long, "loose" clusters while the complete one more compact clusters. Clusters arising from the average linkage procedure are between long chain clusters and tight compact clusters (Everitt et al., 2011).

Determining the number of clusters is an issue of data clustering techniques. In this work, we adopt one of the most common internal cluster validation indexes: the silhouette index (Rousseeuw, 1987). It evaluates the clustering performance based on the pairwise difference of between and within-cluster distances. The optimal cluster number is that maximizes the index value, cutting the dendogram according to it.

In the next two sections, the proposed clustering method (henceforth, "R1 model") is compared both by a simulation and by an application to real data, with respect to two baseline models: that proposed by Fouedjio (2016) (henceforth, "N1 model") and that who takes into account spatial information using the geographical coordinates as additional variables (henceforth, "D1 model").

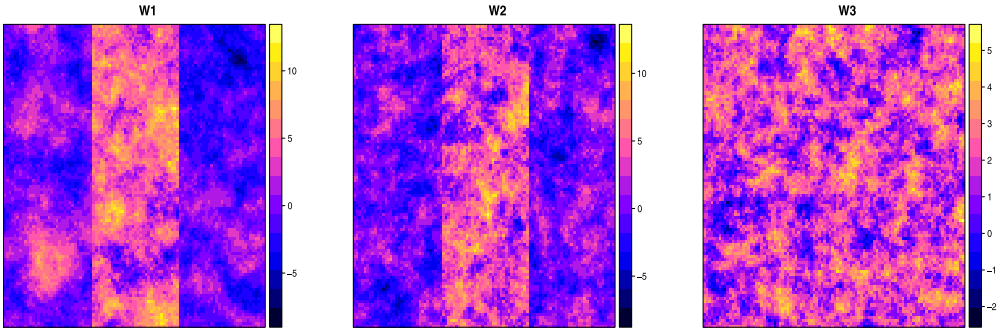
3. Simulation study

3.1. Simulation plan

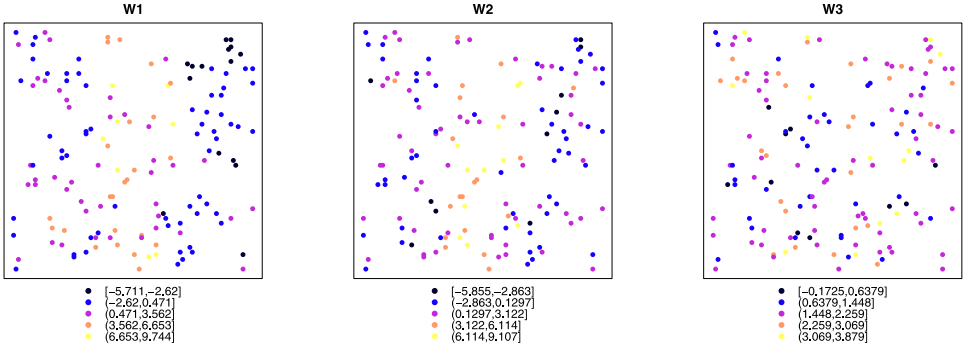
According to the simulation plan of Fouedjio (2016), we generate a clustered bivariate Gaussian random field whose bivariate covariance structure is that of a bivariate Matérn Model as proposed by Gneiting et al. (2010) and implemented in R package `RandomFields` (Schlather et al., 2015).

It is a multivariate stationary isotropic covariance model whose corresponding covariance function only depends on the distance, $\mathbf{h} \geq 0$, between two points. Properly, each marginal covariance function is of the Matérn type as:

$$C_{ii}(\mathbf{h}) = \sigma_i^2 M(\mathbf{h}|\nu_i, a_i) \quad for \quad i = 1, 2 \tag{5}$$



(a) Data generating model



(b) Sample locations

Fig. 4. Example of data simulation and sampling locations.

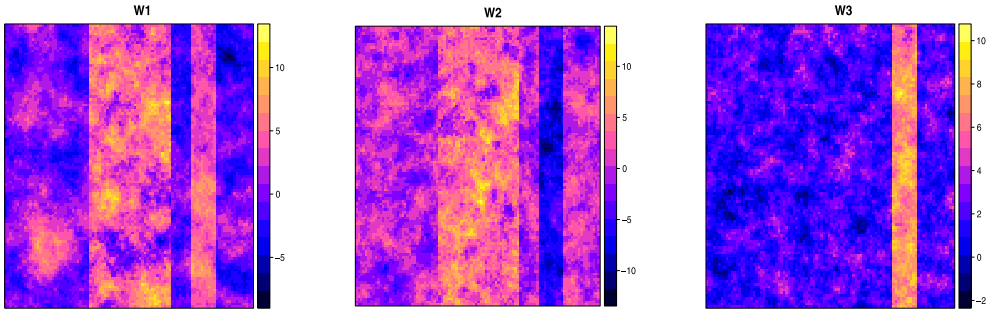
with variance parameter $\sigma_i^2 > 0$, smoothness parameter $\nu_i > 0$ and scale parameter $a_i > 0$. $M(\mathbf{h}|\nu_i, a_i)$ is the spatial correlation at distance $\|\mathbf{h}\|$, as defined by Gneiting et al. (2010). The cross covariance function is also a Matérn function as:

$$C_{ij}(\mathbf{h}) = C_{ji}(\mathbf{h}) = \rho_{ij}\sigma_i^2\sigma_j^2 M(\mathbf{h}|\nu_{ij}, a_{ij}) \quad for \quad i, j = 1, 2 \quad and \quad i \neq j \tag{6}$$

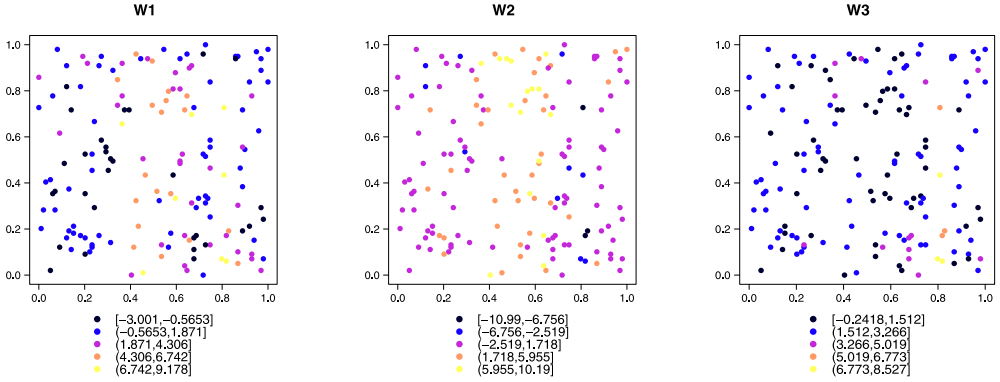
with colocated correlation coefficient ρ_{ij}, smoothness parameter ν_{ij}, and scale parameter a_{ij}.

The clustered bivariate gaussian random field reflects the following clustering structure and parameters: in the subdomains $[0, 1/3] \times [0, 1]$ and $[2/3, 1] \times [0, 1]$, we generate the bivariate gaussian random field $\mathbf{W} = (W_1, W_2)$, with zero mean vector $\boldsymbol{\mu} = (0, 0)$ and Matérn Model covariance parameters: $\rho_{ij} = 0.7$, $\nu_1 = \nu_2 = \nu_{12} = 0.5$, $\boldsymbol{\sigma^2} = (4, 4)$, $a_1 = 0.09$, $a_2 = 0.05$ and $a_{12} = 0.07$. In the subdomains $[1/3, 2/3] \times [0, 1]$, we generate the bivariate gaussian random field with mean vector equal to $\boldsymbol{\mu} = (4, 4)$ and Matérn Model covariance parameters: $\rho_{ij} = 0.7$, $\nu_1 = \nu_2 = \nu_{12} = 0.5$, $\boldsymbol{\sigma^2} = (6, 6)$, $a_1 = 0.05$, $a_2 = 0.03$ and $a_{12} = 0.04$. This leads to two spatial clusters, one of which is non-contiguous. A third variable W_3 is generated according to a Gaussian stationary univariate random function defined on the domain $[0, 1]^2$ with mean $\mu = 2$, $\sigma^2 = 1$ and Matérn stationary correlation parameters of smoothness $\nu = 0.05$ and scale $a = 0.03$, respectively.

According to the above procedure, we simulate 100 independent realizations and, for each one, we randomly sample 50 data locations from each subdomain, resulting in a data set of 150 observations. An example of data simulation and sampling locations is reported in Fig. 4. In order to assess the performance of clustering procedure in presence of outliers, two different simulation scenarios are taken into account, according to the spatial location of outliers, as described in the next section.



(a) Data generating model with outliers



(b) Sample locations with outliers

Fig. 5. Example of data simulation and sampling locations with outliers in the "vertical" area.

3.1.1. Simulation with outliers

The outliers, represented by the vertical area in the subdomain $[0.75, 0.85] \times [0, 1]$, as shown in Fig. 5(a), are generated by performing the following linear transformation to data locations in the subdomain $[0.75, 0.85] \times [0, 1]$:

$$W_{out,i} = W_i + D \quad for \quad i = 1, 2, 3 \tag{7}$$

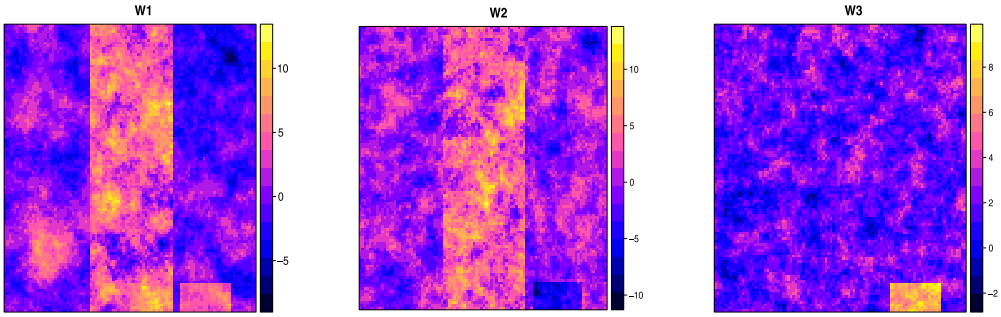
where

$$D = \begin{cases} Q_{3,W_i} + 2 \cdot IQR(W_i) & i = 1, 3 \\ Q_{1,W_i} - 2 \cdot IQR(W_i) & i = 2 \end{cases}$$

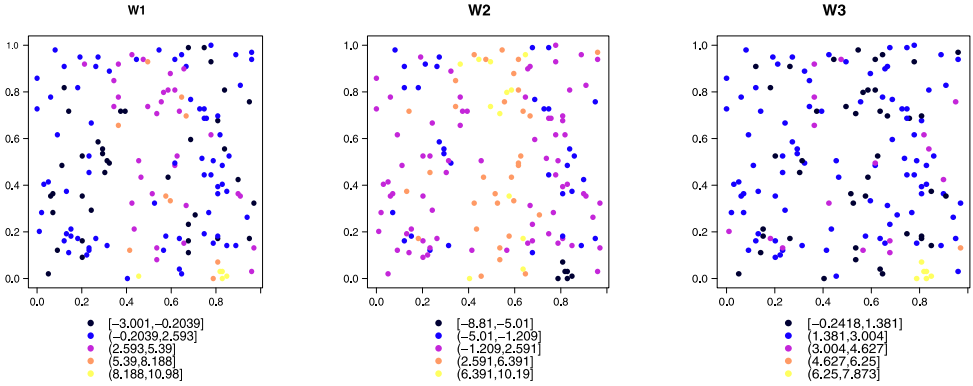
and Q_{1,W_i}, Q_{3,W_i}, $IQR(W_i)$ are the first quartile, the third quartile and the interquartile range of the variable W_i in the subdomain $[2/3, 1] \times [0, 1]$, respectively. The D's definition is based on the well known practice in Statistics that sets as outlier cutoff 1.5 times the interquartile range. The cutoff set at 3 times the interquartile range identifies the so called "extreme values" (Tukey, 1977; Emerson and Strenio, 1983; D'Urso et al., 2011).

The outliers, represented by the rectangular area in the subdomain $[0.70, 0.90] \times [0, 0.1]$, as shown in Fig. 6(a), are generated by applying the same linear trasformation in Eq. (7) to data locations in the subdomain $[0.70, 0.90] \times [0, 0.1]$.

Therefore, for each scenario, we simulate 100 independent realizations by randomly sampling 50 data locations from each subdomain (excluding outliers). From the subdomain interested by the outliers, we sample $n = 6$ and $n = 12$ data locations, respectively. Therefore, we generate 400 data



(a) Data generating model with outliers



(b) Sample locations

Fig. 6. Example of data simulation and sampling locations with outliers in the "rectangular" area.

sets: 100 data sets of 156 observations and 100 data sets of 162 observations, replicated for the two scenarios. An example of sampling locations with outliers is reported in Fig. 5(b) and Fig. 6(b), respectively.

All comparisons are carried out by means of the Rand index (Rand, 1971), that is a measure of the similarity between two data clusterings: the closer the index is to 1, the better is the obtained clustering.

In the next paragraphs, we compare the performance of the R1 model with respect to two baseline models: the N1 and D1 ones. As described in the paragraphs 2.2 and 2.3, we remember that the N1 and R1 models are based on the same kernel estimator, they differ with respect to the bandwidth computation and to the exponential transformation applied to the original dissimilarity index.

3.2. Simulation results

In this section, we analyse the algorithm performances with reference to the average and complete linkage only since, according to Fouedjio (2016), the single linkage method has a worst performance in this specific context.

The preliminary phase of the analysis takes into account data simulation without outliers. Table 1 reports the Rand index mean values for the D1, N1 and R1 models respectively, while Fig. 7 shows the corresponding boxplots, confirming no remarkable differences between the R1 and N1 models

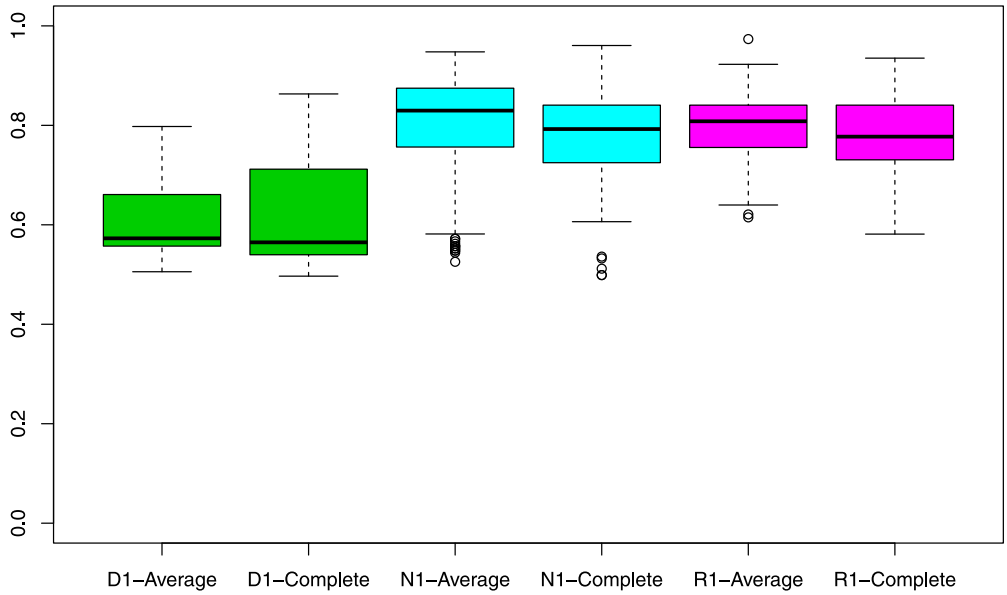


Fig. 7. Simulation results without outliers: boxplots of Rand index.

Table 1
Simulation results without outliers: means of Rand index.

Method	D1 model	N1 model	R1 model
Average linkage	0.605	0.792	0.799
Complete linkage	0.617	0.776	0.776

Table 2
Scenario with outliers in the vertical area: means of Rand index.

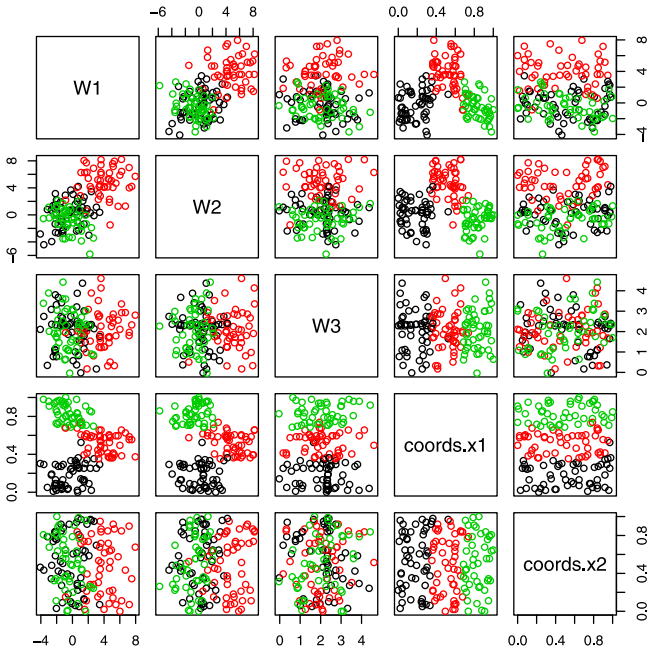
	6 outliers			12 outliers		
	D1 model	N1 model	R1 model	D1 model	N1 model	R1 model
Average linkage	0.553	0.551	0.793	0.553	0.541	0.807
Complete linkage	0.553	0.562	0.781	0.554	0.524	0.785

when data are not affected by anomalies. On the contrary, the D1 model, frequently used in the applications, shows a very low performance with respect to based kernel estimator models.

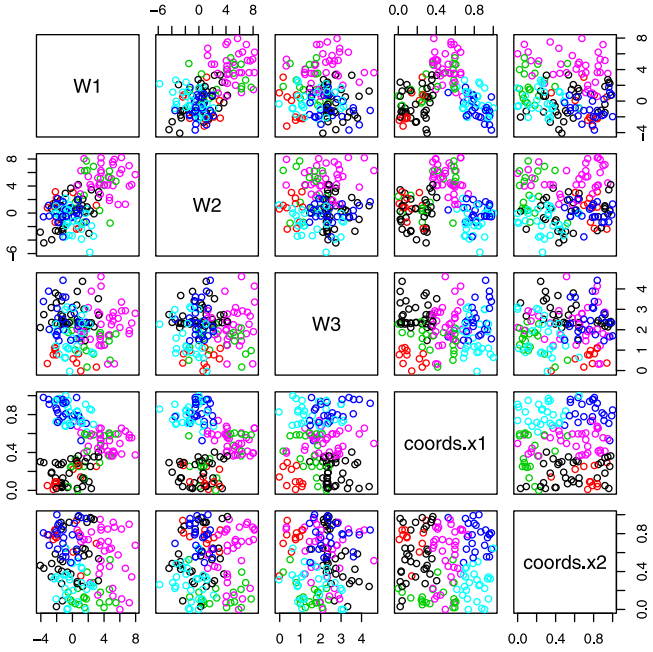
To reinforce all these considerations, we can also inspect the scatterplots in Figs. 8, 9, 10. The D1 model is absolutely not be able to reproduce the two non-contiguous spatially clusters with a tendency to identify many more clusters than they should be.

3.2.1. Scenarios with outliers

The first scenario refers to outliers represented by the vertical area in the subdomain $[0.75, 0.85] \times [0, 1]$ as shown in Fig. 5. By inspecting the boxplots of the Rand index (Fig. 11) as well as its average values (Table 2), the better performance of the R1 model, with respect to the baseline ones, becomes strongly evident. The scatterplots in Figs. 12, 13, 14, referred to the case of six outliers, definitely point out the marked difference among the three models. In fact, different than the baseline models, our proposal is able, at the same time, to recover the expected spatial structure of data (the two non-contiguous groups) and to detect spatial outliers (the green points in the

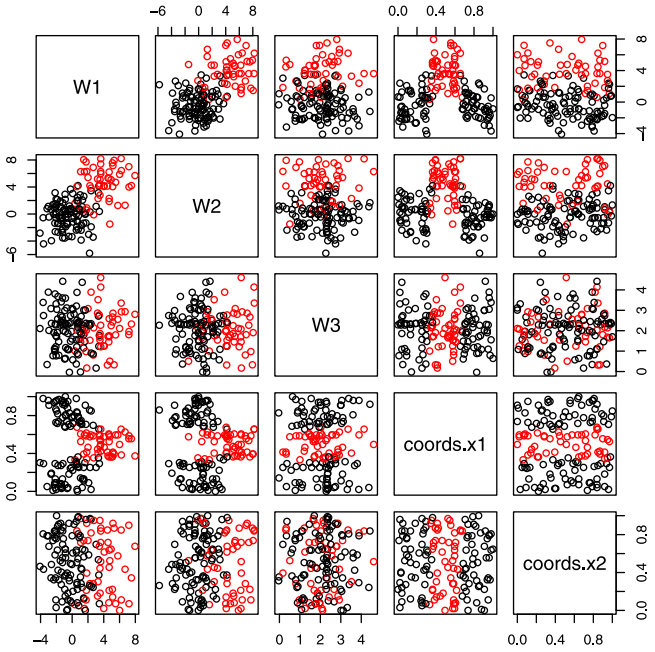


(a) Average linkage partition

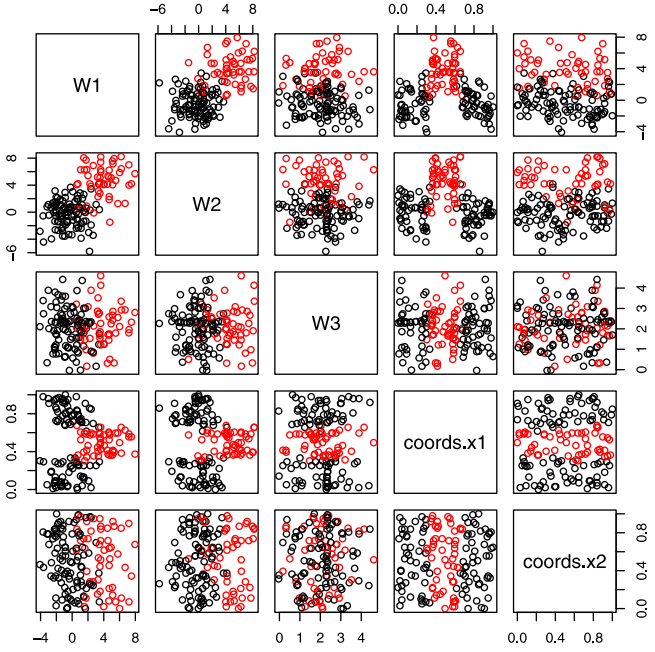


(b) Complete linkage partition

Fig. 8. Scenario without outliers: D1 model.

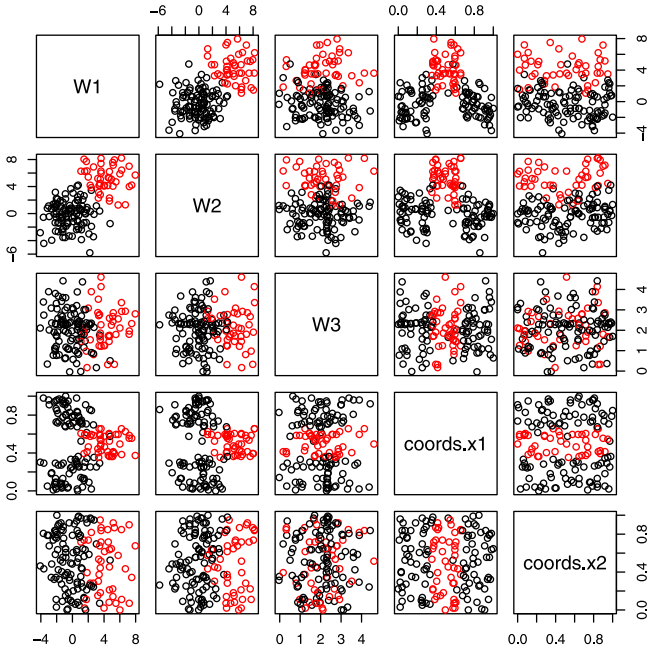


(a) Average linkage partition

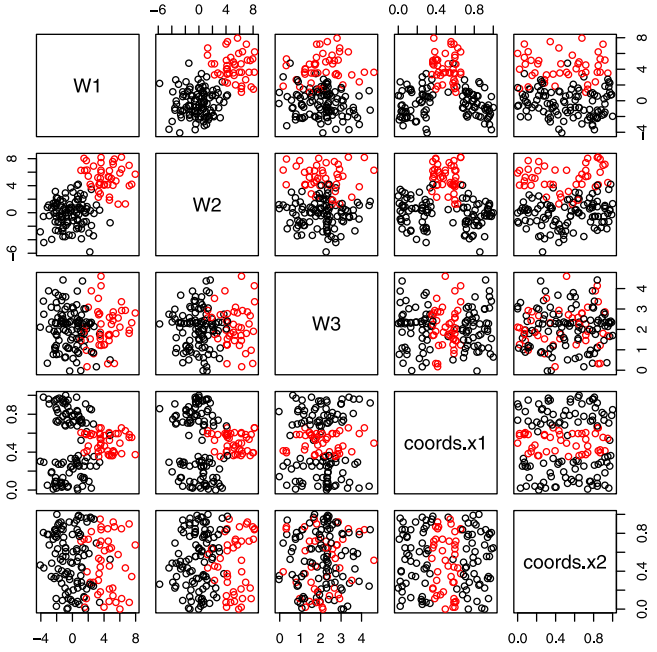


(b) Complete linkage partition

Fig. 9. Scenario without outliers: N1 model.

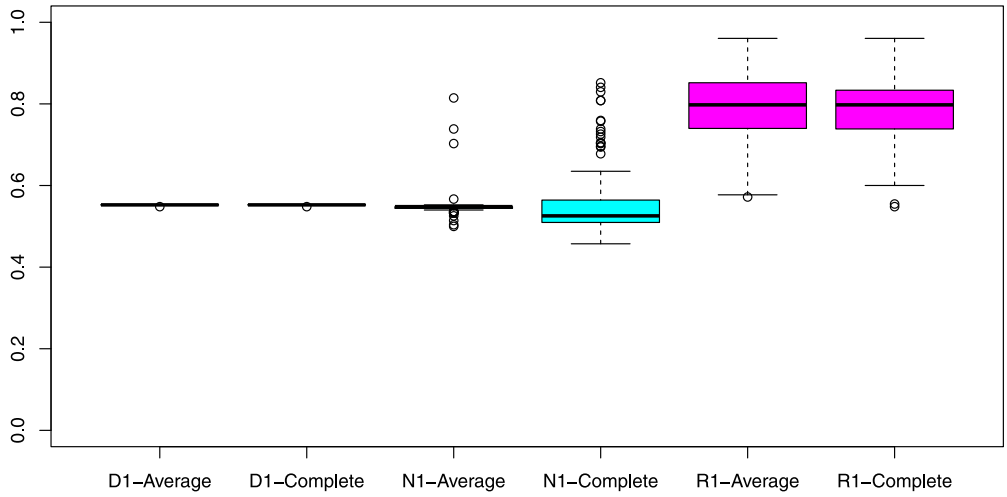


(a) Average linkage partition

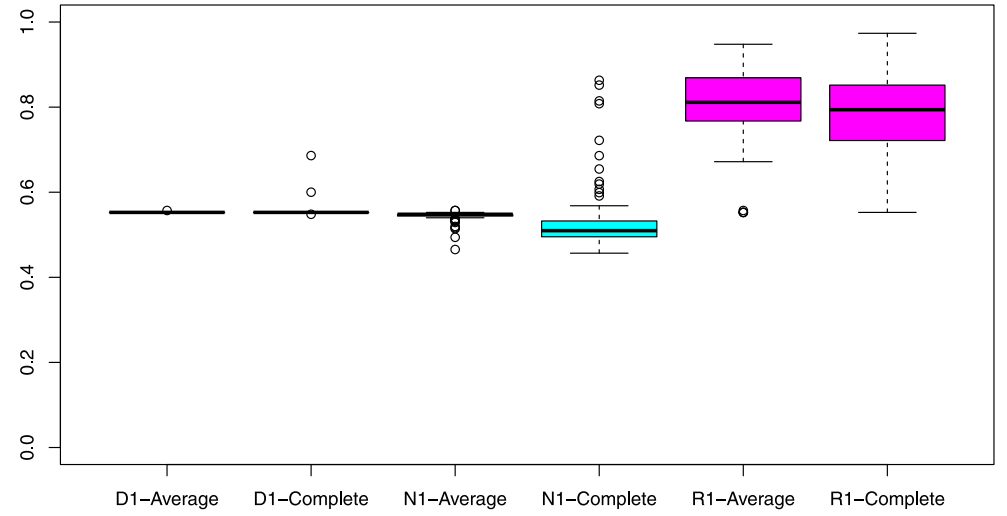


(b) Complete linkage partition

Fig. 10. Scenario without outliers: R1 model.



(a) 6 outliers

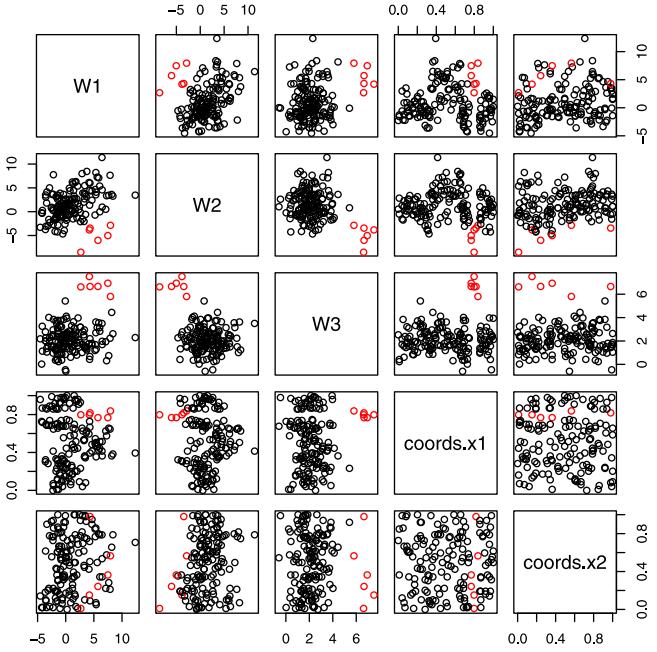


(b) 12 outliers

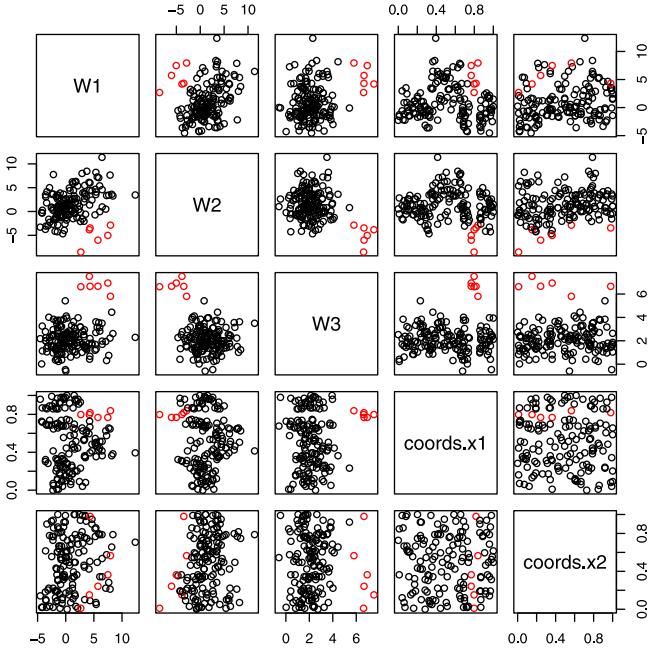
Fig. 11. Scenario with outliers in the vertical area: boxplot of Rand index.

scatterplot of Fig. 14). The scatterplots in Figs. 12 and 13 clearly show the disruptive effect of the outliers for the baseline models: there is an evident loss of information about the expected spatial data structure, that is characterized by two non-contiguous groups. The D1 and N1 models, indeed, assign data points to the same cluster.

The second scenario refers to outliers represented by the rectangular area in the subdomain $[0.70, 0.90] \times [0, 0.1]$ as shown in Fig. 7. By inspecting the boxplots of the Rand index (Fig. 15) as well as its average values (Table 3), the better performance of the R1 model, with respect to the baseline ones, is strongly remarkable also in this case. The scatterplots in Figs. 16, 17, 18, referred to the case of six outliers, provide the same evidences of the first scenario.

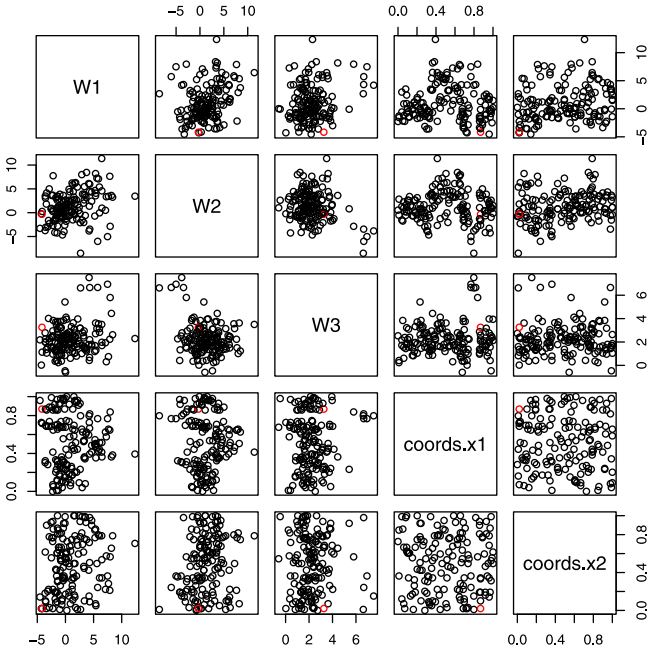


(a) Average linkage partition

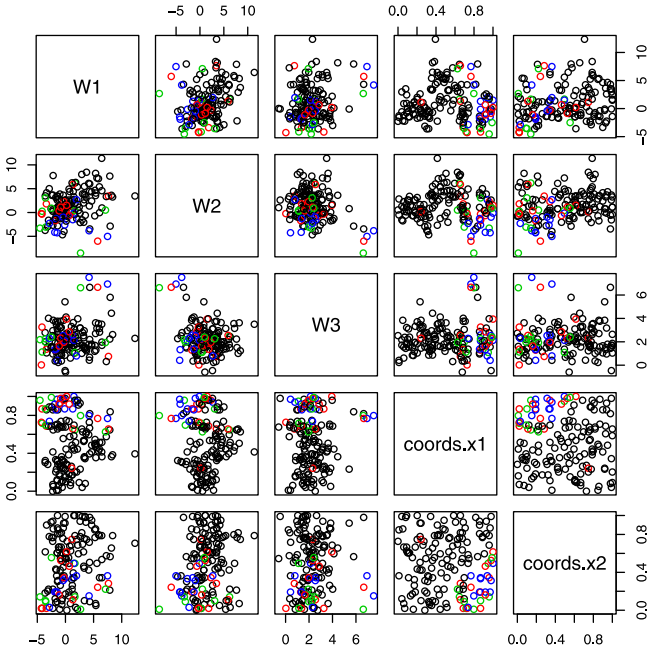


(b) Complete linkage partition

Fig. 12. Scenario with 6 outliers in the vertical area: D1 model.

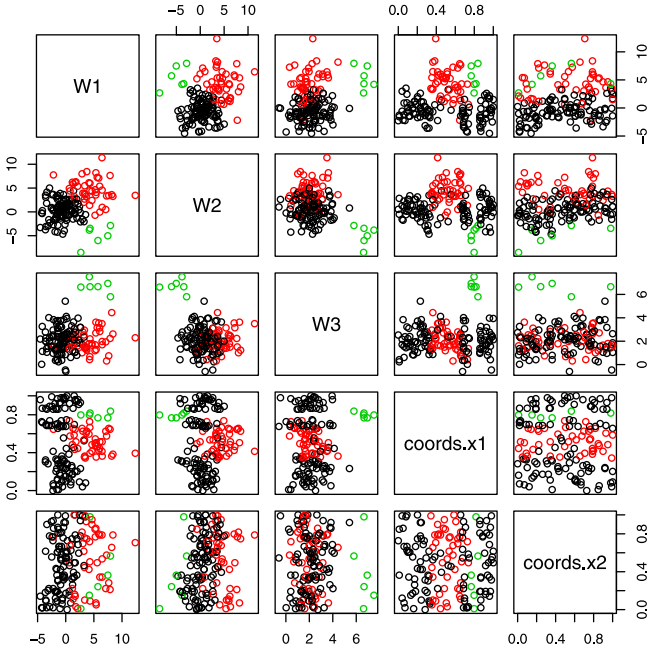


(a) Average linkage partition

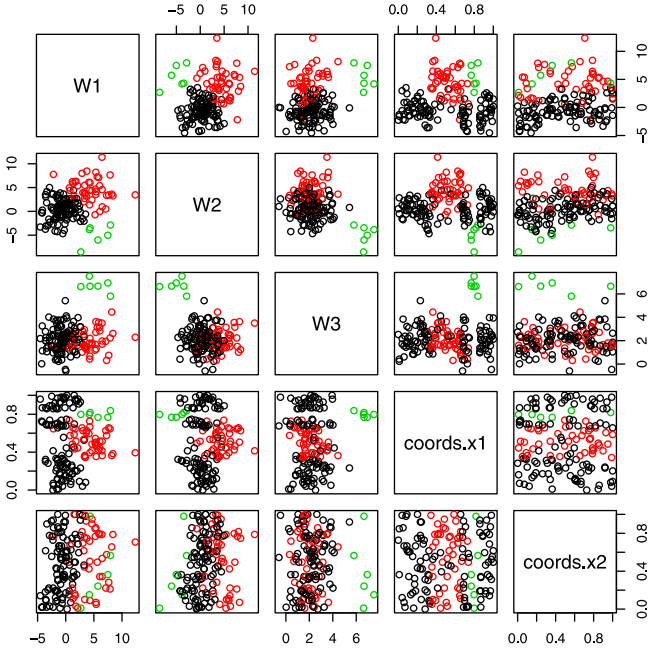


(b) Complete linkage partition

Fig. 13. Scenario with 6 outliers in the vertical area: N1 model.

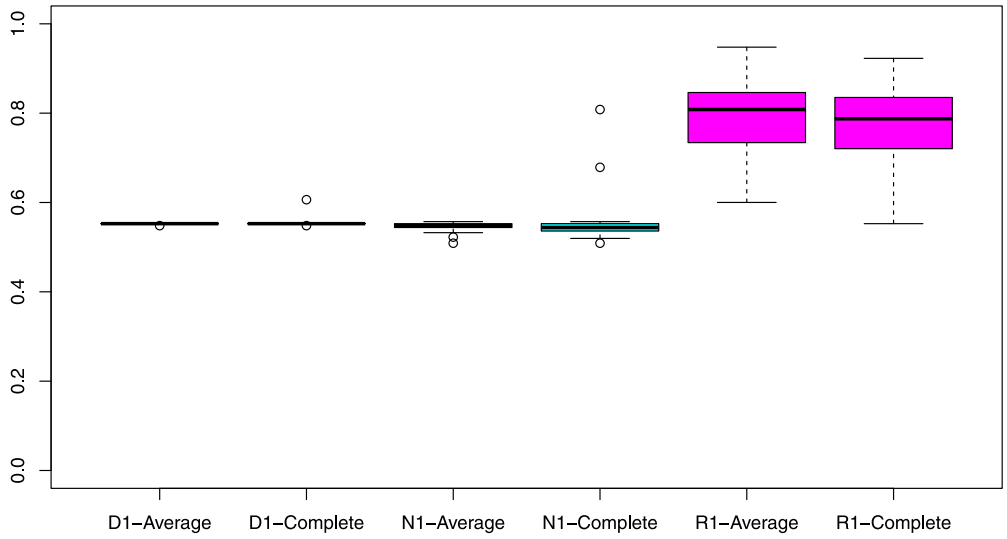


(a) Average linkage partition

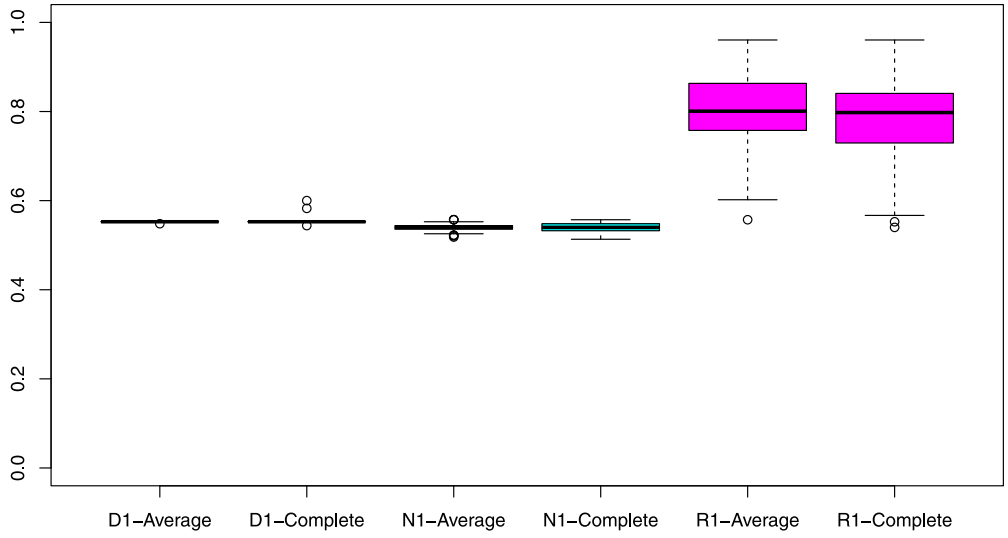


(b) Complete linkage partition

Fig. 14. Scenario with 6 outliers in the vertical area: R1 model.



(a) 6 outliers

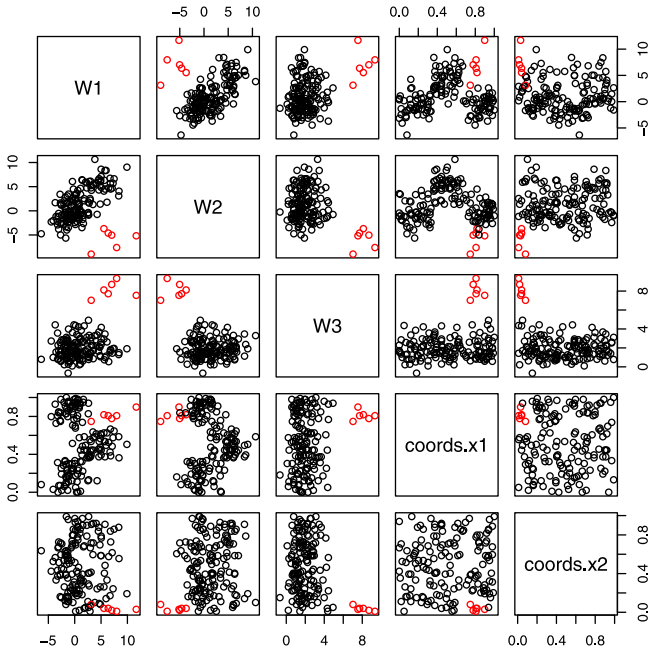


(b) 12 outliers

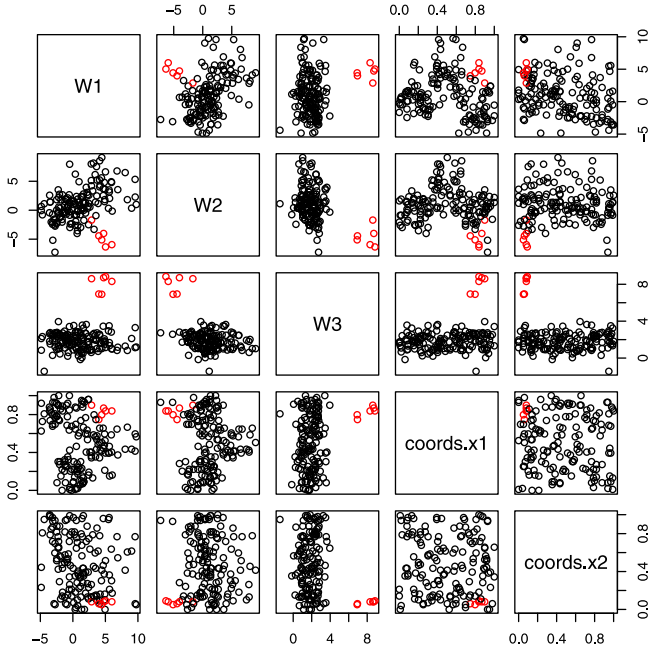
Fig. 15. Scenario with outliers in the rectangular area: boxplot of Rand index.

We do not report the scatterplots referred to the case of twelve outliers because they lead to the same conclusions for both cases.

Simulation results confirm the better performance of our proposal, based on a slightly different computation of kernel bandwidth and on an exponential transformation of the dissimilarity index, able to accomplish the following two issues: to guarantee spatial contiguity of clusters and to detect outliers in the attribute space ensuring that the underlying groups structure is preserved.

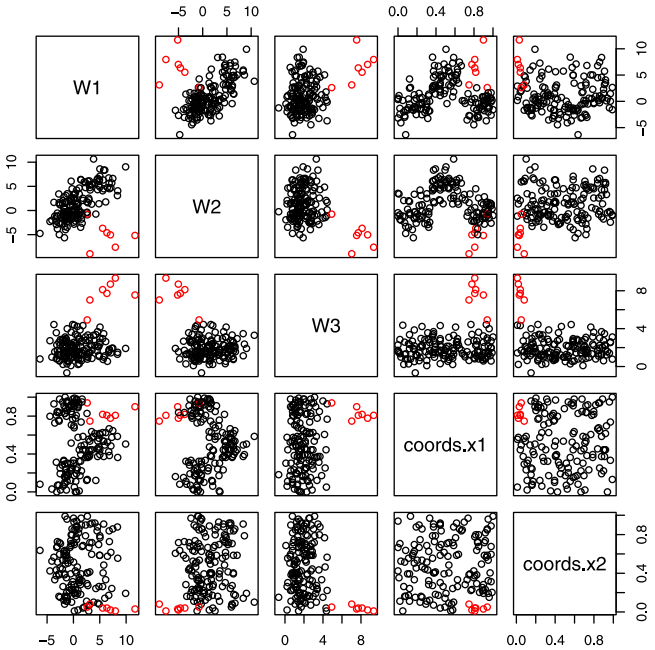


(a) Average linkage partition

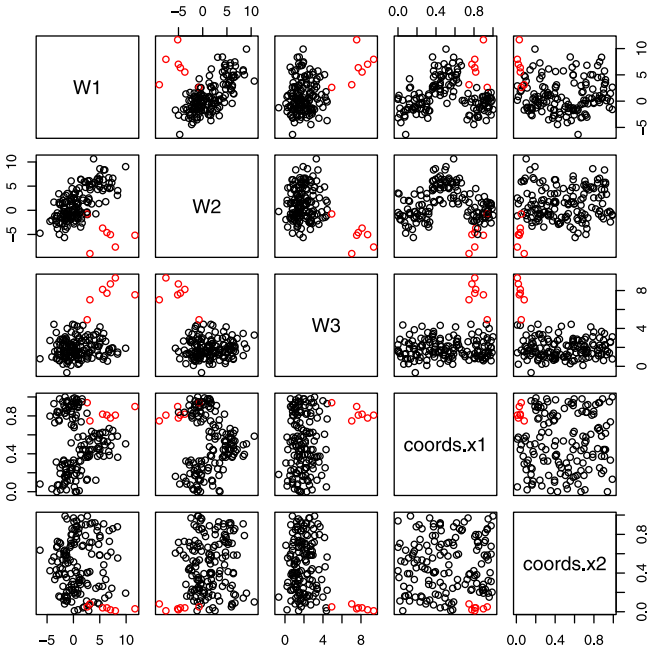


(b) Complete linkage partition

Fig. 16. Scenario with 6 outliers in the rectangular area: D1 model.

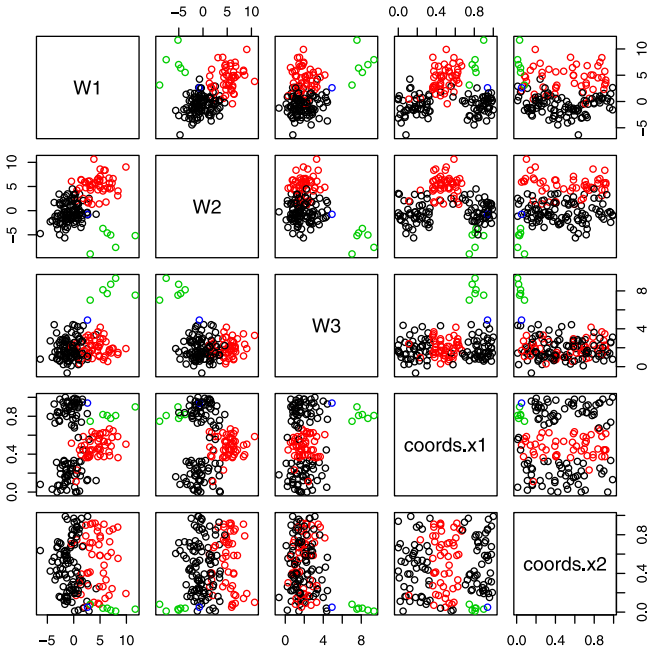


(a) Average linkage partition

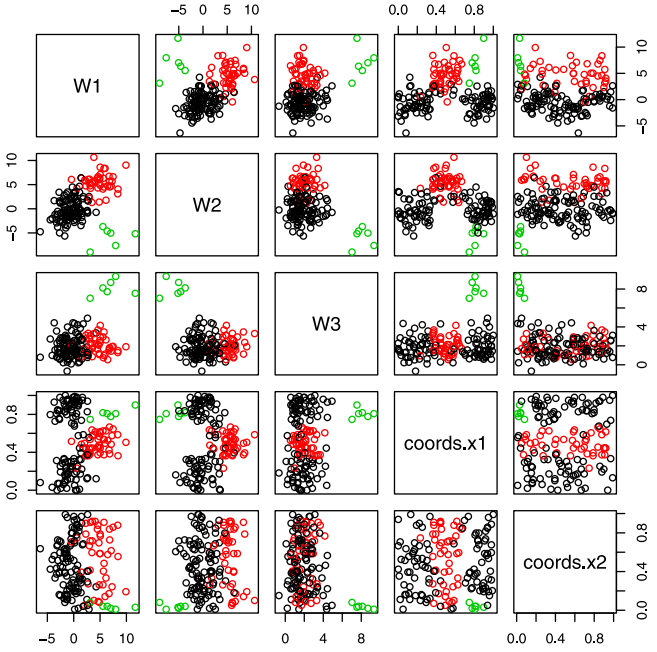


(b) Complete linkage partition

Fig. 17. Scenario with 6 outliers in the rectangular area: N1 model.



(a) Average linkage partition



(b) Complete linkage partition

Fig. 18. Scenario with 6 outliers in the rectangular area: R1 model.

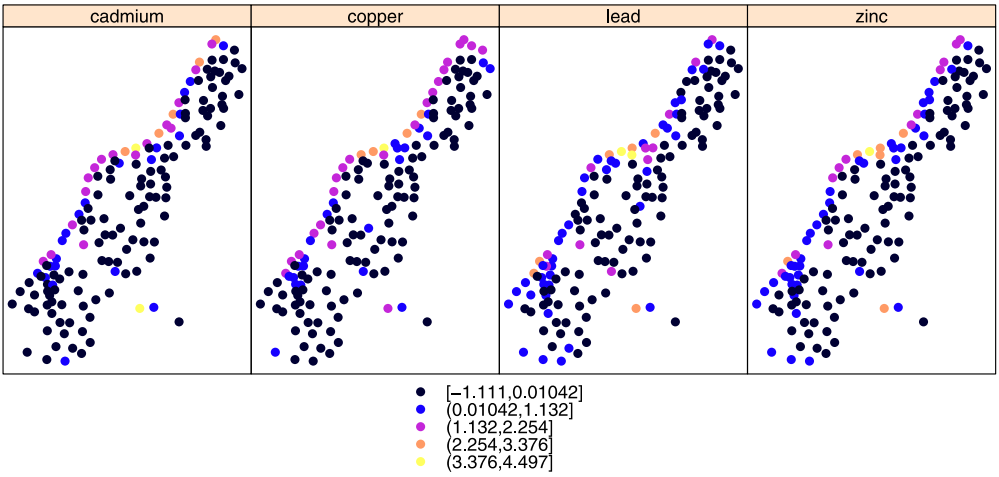


Fig. 19. Meuse data set: the standardized topsoil concentrations of heavy metals.

Table 3
Scenario with outliers in the rectangular area: means of Rand index.

	6 outliers			12 outliers		
	D1 model	N1 model	R1 model	D1 model	N1 model	R1 model
Average linkage	0.553	0.547	0.793	0.553	0.540	0.801
Complete linkage	0.553	0.547	0.770	0.553	0.539	0.787

4. Application to real data

4.1. Meuse data set

In this section, the proposed clustering method and the baseline ones are applied to a georeferenced data set available in the R package `sp` (Pebesma and Bivand, 2005), including locations and topsoil concentrations (ppm) of cadmium, copper, lead and zinc metals, collected in a flood plain of the river Meuse, near the village of Stein (NL). Heavy metal concentrations are from composite samples of an area of approximately 15 m $\times$ 15 m.[1] The sample data location are 155 and all variables and coordinates[2] have been standardized.

From Fig. 19, we can see that the river bank is characterized by high topsoil concentrations of heavy metals. The presence of several outliers (in terms of very high heavy metals concentrations) could be suggested by a preliminary visual inspection of the plots of Fig. 20. Moreover, the visual inspection of the multivariate data via a Sammon projection in two dimensions using the Euclidean metric (see Fig. 21), also confirms the presence of several data points which lie far from the bulk of data. Clustering results for all the three models, with reference to the average and complete linkage respectively, are reported in Fig. 23. As suggested by simulation study, the baseline models produce the same unsatisfactory results since they are highly affected by the presence of outliers. Identifying only two clusters, they are not be able to take into account the underlying structure of data, for which the main differences, in the attribute and non attribute space, depend on the proximity of the location wit respect to the river bank. These features are taken into account when we apply

[1] Field data were collected by Ruud van Rijn and Mathieu Rikken, were compiled for R by Edzer Pebesma and the description was extended by David Rossiter.

[2] The coordinates have not been standardized for the N1 model.

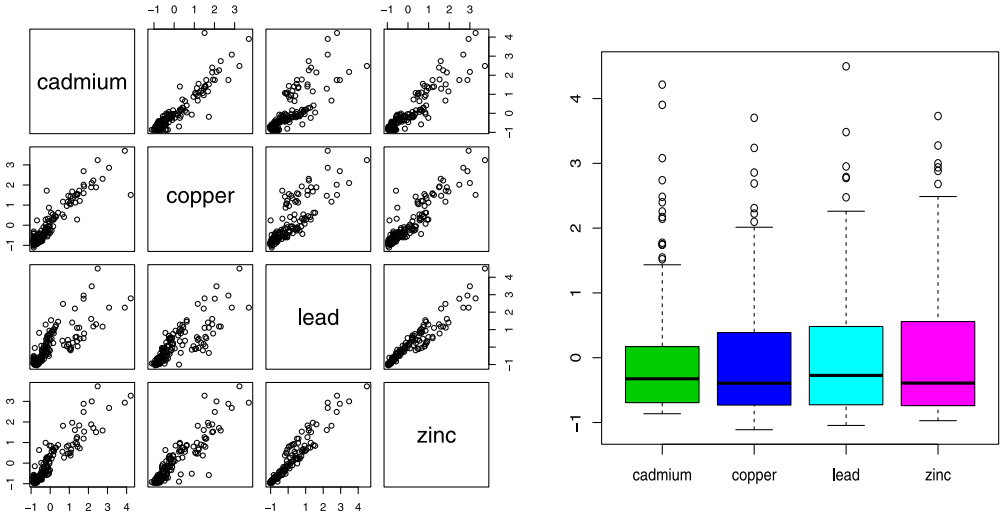


Fig. 20. Meuse data set: scatterplot and boxplots of the standardized topsoil concentrations of heavy metals.

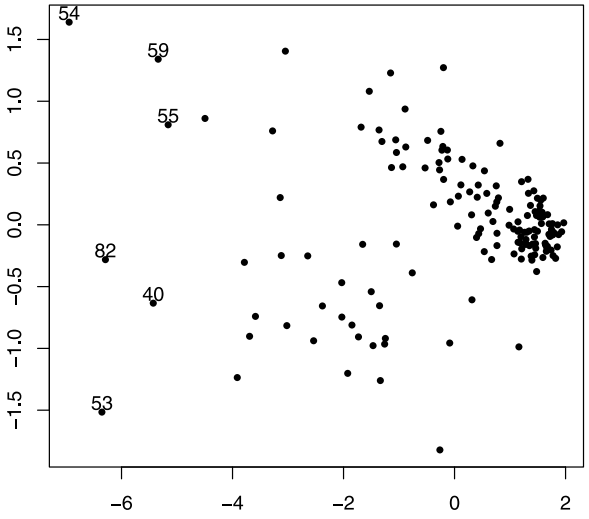
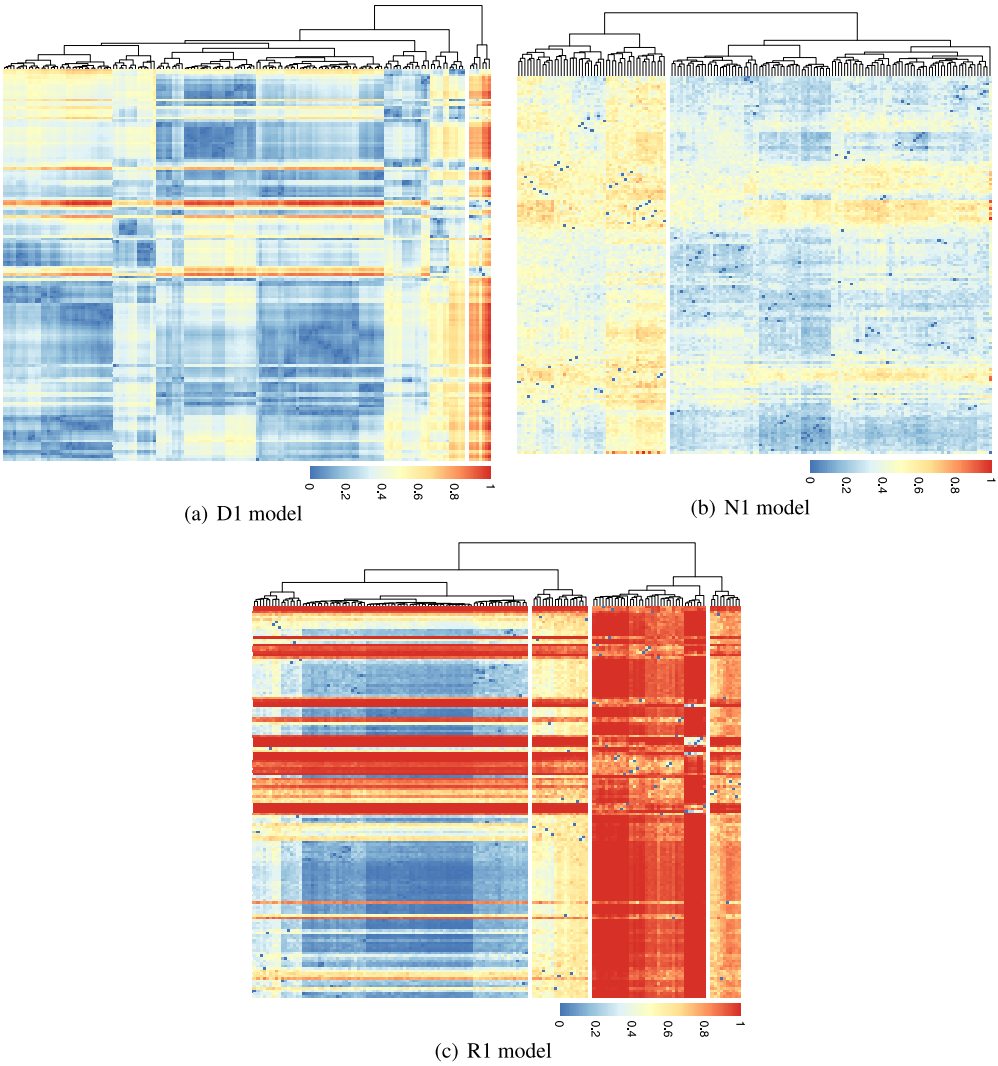


Fig. 21. Meuse data set: Sammon projection in two dimensions of the standardized topsoil concentrations of heavy metals.

the R1 model. This method is able, at the same time, to detect the outliers group (yellow points, corresponding to those numbered in Fig. 21) and to recover the underlying structure of spatial data. It seems to balance the main two issues concerning spatial contiguity and proximity in the attribute space. The results are confirmed by the dendograms shown in Fig. 22.

In Table 4, we report the mean and standard deviation values of the four clusters identified by the robust R1 method, with complete linkage. The third group identifies the less pollutant zone, followed by the second one, while the fourth one includes all locations with heavy metal concentrations out of control. The first group identifies the second high pollutant zone.

In addition, we consider the logarithmic transformation in the comparative study. As we can see, the role of outliers is strongly reduced (see Figs. 24 and 25) and the resulting group structure



(a) D1 model (b) N1 model

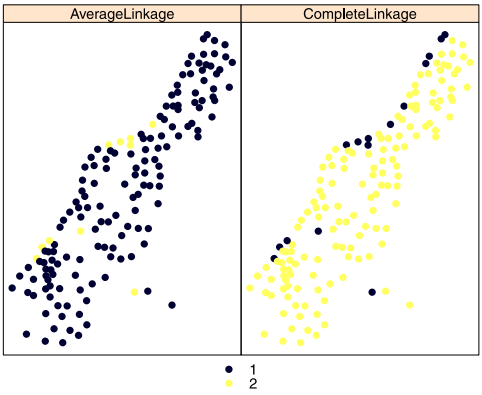
(c) R1 model

Fig. 22. Meuse data set: dendograms using the complete linkage (standardized variables).

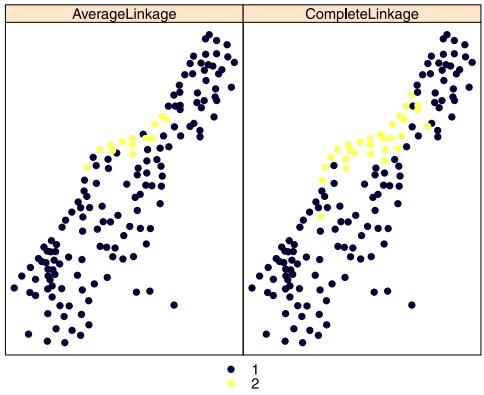
Table 4
Meuse data set: means and standard deviations of raw variables for each spatial cluster (complete linkage).

	Group 1		Group 2		Group 3		Group 4	
	Mean	sd	Mean	sd	Mean	sd	Mean	sd
Cadmium	9.269	1.994	5.311	1.970	1.411	0.987	13.583	3.451
Copper	82.077	7.566	57.964	12.035	27.185	7.361	103.833	18.659
Lead	309.462	77.760	215.464	59.192	99.667	50.516	491.833	94.516
Zinc	1078.385	131.090	697.679	116.644	274.444	136.810	1602.000	135.887

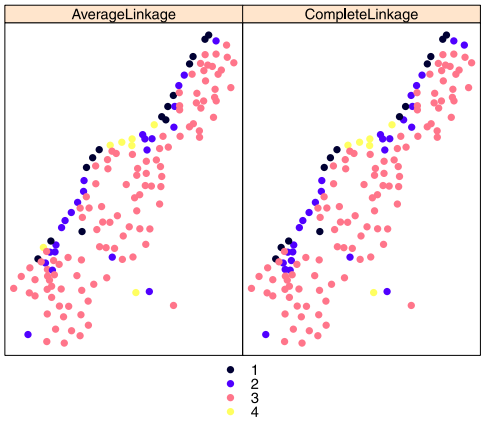
is shown in Fig. 26 for all three models. Also in this situation, the R1 model (complete linkage) seems more able to distinguish between the different pollutant zones, preserving spatial contiguity



(a) D1 model



(b) N1 model



(c) R1 model

Fig. 23. Meuse data set: clustering results (standardized variables).

but also detecting non-contiguous locations with very high metals concentrations. The clustering structure recovered by the D1 model, with reference to the average linkage, is not surprising since, in the simulation study, we have already pointed out its tendency to identify more groups than the existing ones. The results are confirmed by the dendograms shown in Fig. 27.

4.2. Jura data set

In this section, the proposed clustering method and the baseline ones are applied to a geo-referenced data set available in the R package `gstat` (Pebesma, 2004), named "Jura data set", including locations and topsoil concentrations of the following heavy metals: cadmium, cobalt, chromium, copper, lead, nickel and zinc. Georeferencing was based on two control points in the Swiss grid system.[3] The sample data locations are 100 and all variables and coordinates[4] have been standardized. The plots of Figs. 28 and 29 suggest the possible presence of multivariate outliers in terms of very high heavy metals concentrations (two points, in particular). This is also confirmed

[3] Georeferencing is due to David Rossiter while data comes from Pierre Goovaerts' book (Goovaerts, 1997).

[4] The coordinates have not been standardized for the N1 model.

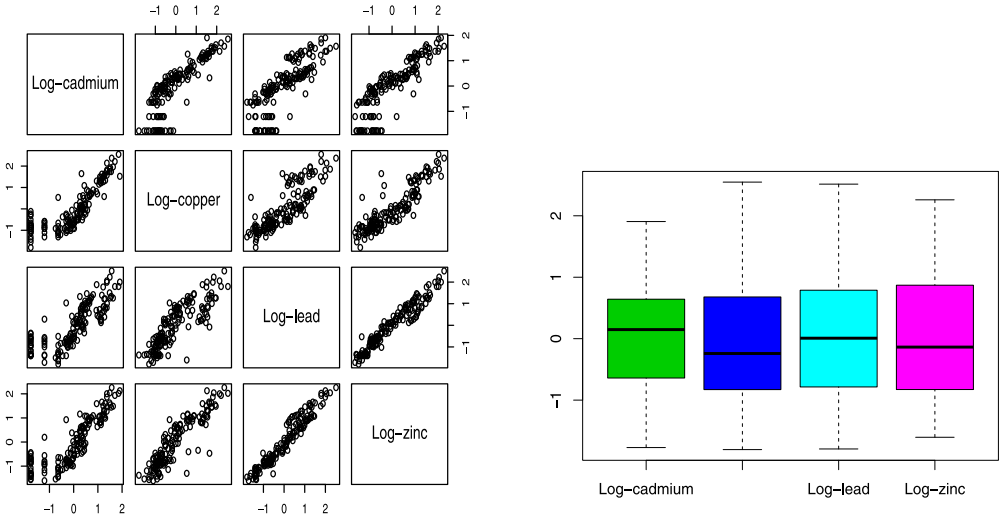


Fig. 24. Meuse data set: scatterplot and boxplots of the standardized topsoil concentrations of heavy metals with logarithmic transformation.

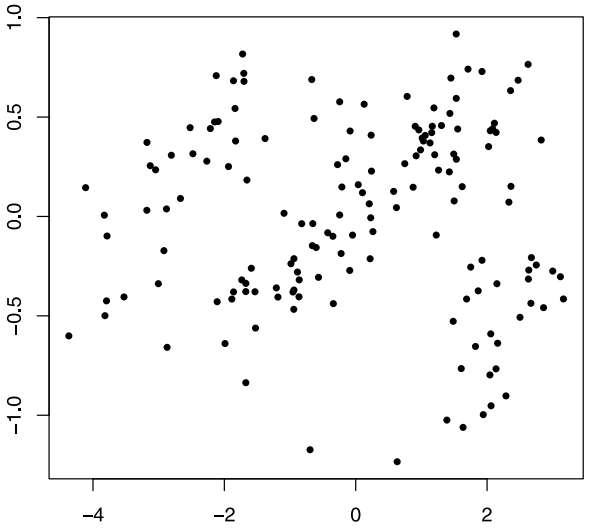
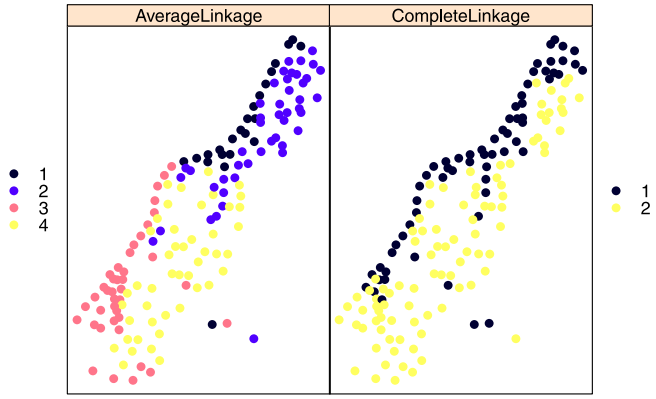


Fig. 25. Meuse data set: Sammon projection in two dimensions of the standardized topsoil concentrations of heavy metals with logarithmic trasformation.

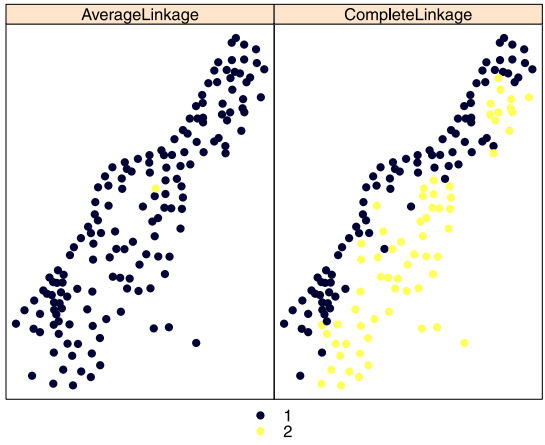
by inspecting the Sammon projection in two dimensions of the same data set using the Euclidean metric, as shown in Fig. 30.

The results of clustering techniques are reported in Fig. 31. Also in this case, the D1 model produces unsatisfactory results being highly affected by the presence of outliers. The N1 model shows a slightly better performance, only for the complete linkage.

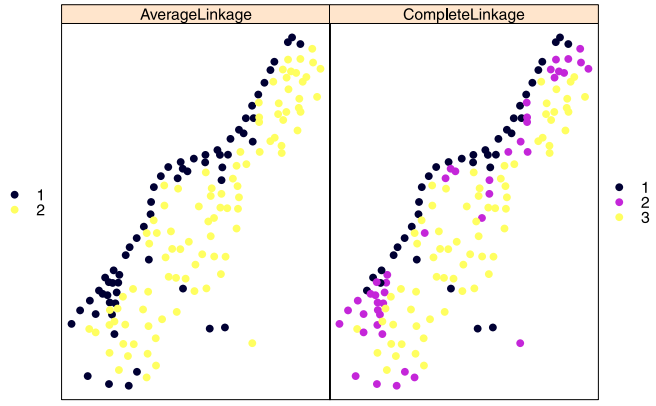
The R1 model confirms its capability both to detect the outliers (the two yellow points, that are the units 30 and 45 identified in the Sammon map of Fig. 30) and to recover the group structure. The results are confirmed by the dendograms shown in Fig. 32.



(a) D1 model

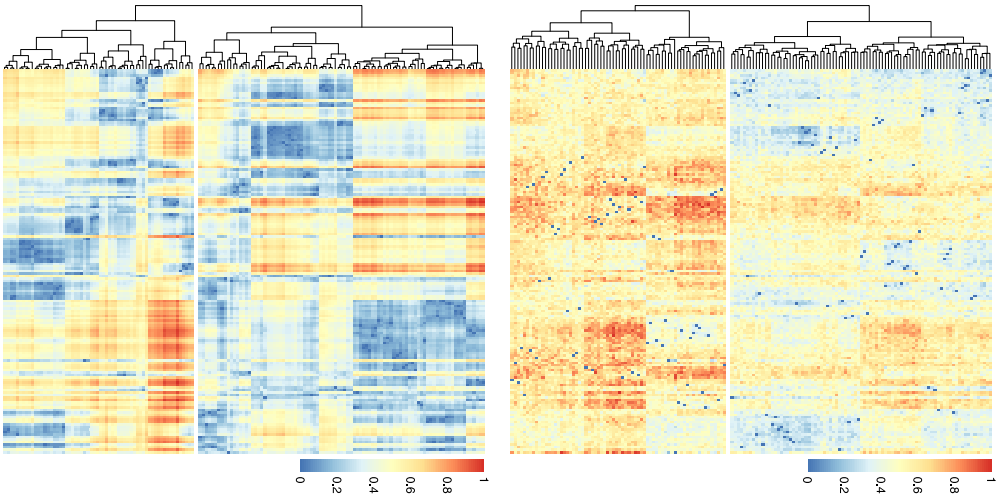


(b) N1 model

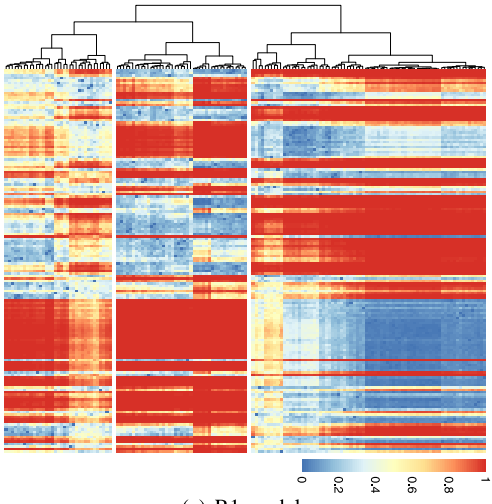


(c) R1 model

Fig. 26. Meuse data set: clustering results (log-transformed standardized variables).



(a) D1 model (b) N1 model



(c) R1 model

Fig. 27. Meuse data set: dendograms using the complete linkage (log-transformed standardized variables).

In Table 5, we report the mean and standard deviation values of the three clusters identified by the robust R1 method, with complete linkage. The group 3 is, clearly, composed of the two yellow outliers, with values out of control for Copper, Lead and Zinc metals. Clustering reveals the presence of two groups, the former characterized by lower concentrations of all heavy metals than the latter.

Following Fouedjio (2017), in this application, we do not consider the log transformation as it modifies the order structure of the raw data.

5. Conclusions

In this paper, we proposed a robust agglomerative hierarchical clustering for multivariate georeferenced data that is a robust version of the clustering method suggested by Fouedjio (2016). In particular, by considering a non-parametric kernel estimator of the direct and cross variograms,

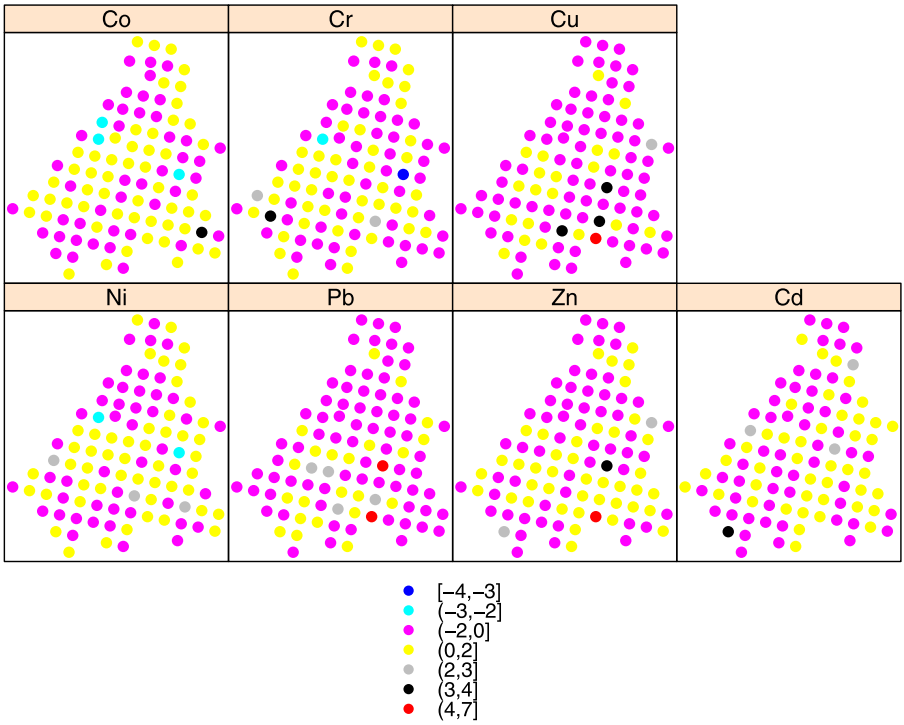


Fig. 28. Jura data set: the standardized topsoil concentrations of heavy metals.

Table 5
Jura data set: means and standard deviations of raw variables for each spatial cluster (complete linkage).

	Group 1		Group 2		Group 3	
	Mean	sd	Mean	sd	Mean	sd
Cadmium	0.938	0.551	1.736	0.641	1.772	0.011
Cobalt	8.554	3.414	11.961	2.738	10.860	0.764
Chromium	29.897	7.371	43.520	7.871	40.760	0.339
Copper	14.105	8.835	32.901	27.349	140.800	19.516
Nickel	16.919	6.156	27.290	5.432	27.660	4.441
Lead	43.994	16.993	66.758	33.573	269.980	42.455
Zinc	59.942	16.127	101.947	23.859	225.920	47.970

we proposed an exponential transformation of the dissimilarity measure for geographical data suggested by Fouedjio (2016). The new dissimilarity measure is robust in the sense that it is capable to neutralize the negative effects of possible spatial outliers in the clustering process. Using the proposed robust dissimilarity measure in an agglomerative hierarchical clustering framework the natural structure of the spatial data is not altered, *i.e.* the natural clusters of spatial units are not affected by the negative influence of possible spatial outliers.

In order to show the usefulness and the performance of the proposed robust clustering method, we compared it with two clustering methods: that suggested by Fouedjio (2016) and that using locations as additional variables. Simulation studies and an application to real data showed very good performance of our clustering method. In future, it will be interesting to investigate the performance of our robust dissimilarity measure for multivariate georeferenced data in other unsupervised clustering methodologies, *i.e.* in a fuzzy framework, and/or in spatial-time domain

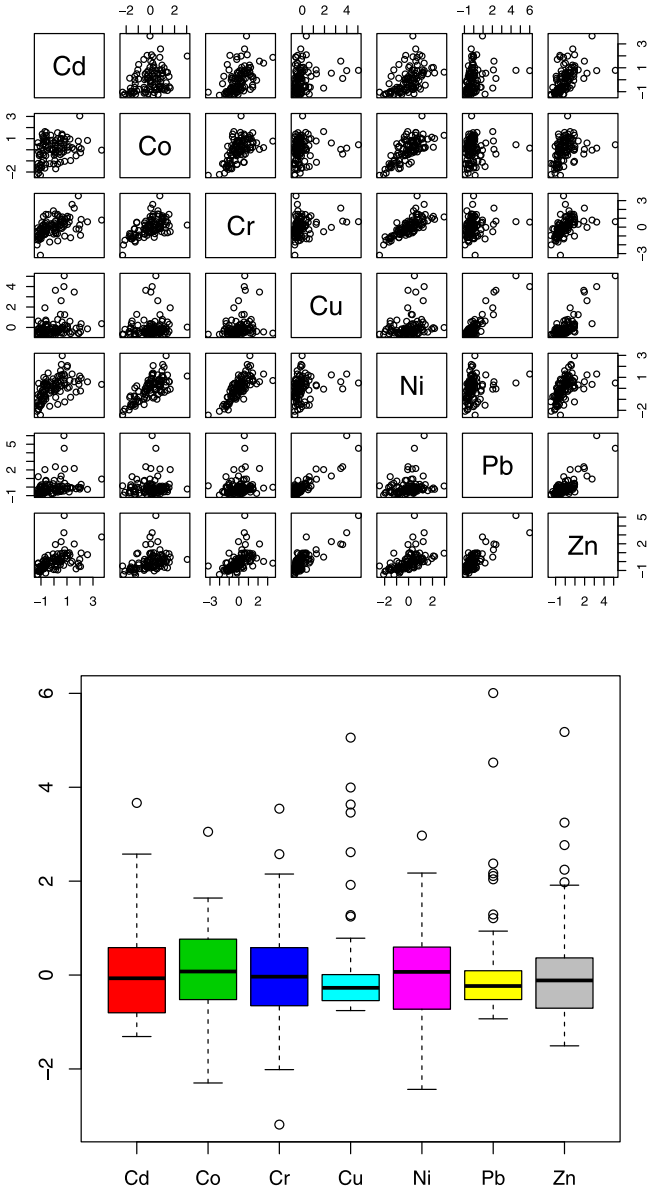


Fig. 29. Jura data set: scatterplot and boxplots of the standardized topsoil concentrations of heavy metals.

(Coppi et al., 2010; Disegna et al., 2017; D'Urso et al., 2019a) with land and soil applications (D'Urso et al., 2019b; Rossiter et al., 2017).

Acknowledgment

The authors thank the Editor and the referees for their useful comments and suggestions which helped to improve the quality and presentation of this manuscript.

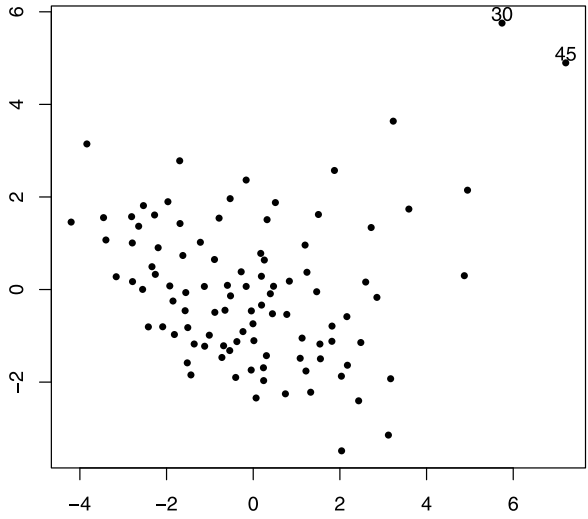
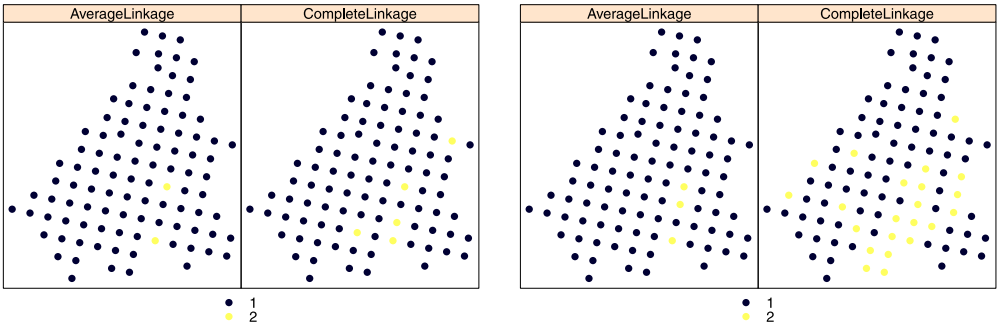
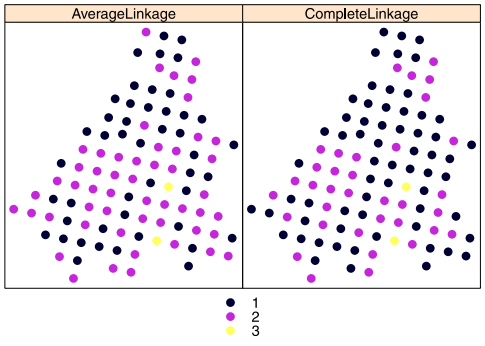


Fig. 30. Jura data set: Sammon projection in two dimensions of the standardized topsoil concentrations of heavy metals.

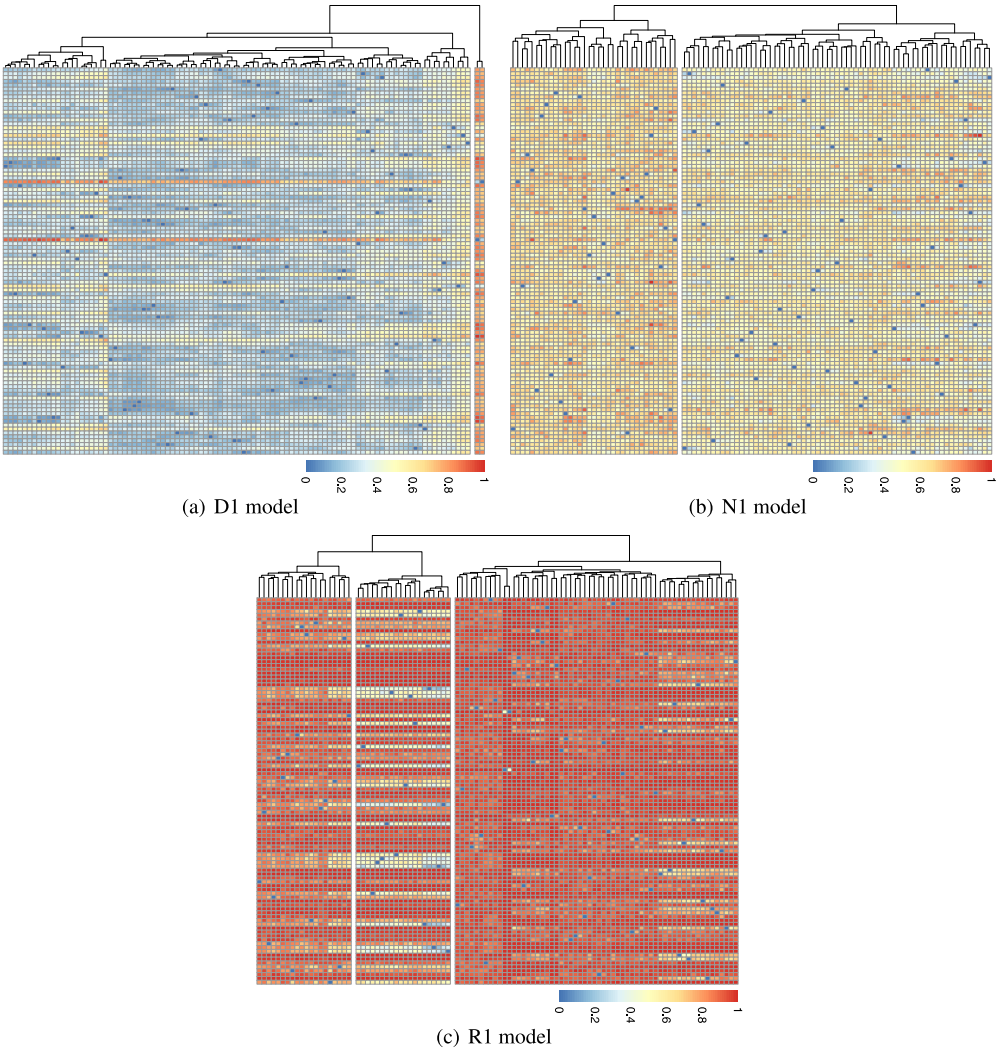


(a) D1 model (b) N1 model



(c) R1 model

Fig. 31. Jura data set: clustering results (standardized variables).



(a) D1 model (b) N1 model

(c) R1 model

Fig. 32. Jura data set: dendograms using the complete linkage (standardized variables).

References

Anselin, L., 1995. Local indicators of spatial association—LISA. Geogr. Anal. 27 (2), 93–115.
Barnett, V., 1978. The study of outliers: purpose and model. J. R. Stat. Soc. Ser. C. Appl. Stat. 27 (3), 242–250.
Barnett, V., Lewis, T., 1994. Outliers in Statistical Data. Wiley.
Cerioli, A., Riani, M., 1999. The ordering of spatial data and the detection of multiple outliers. J. Comput. Graph. Stat. 8 (2), 239–258.
Chen, D., Lu, C.T., Kou, Y., Chen, F., 2008. On detecting spatial outliers. Geoinformatica 12 (4), 455–475.
Coppi, R., D'Urso, P., Giordani, P., 2010. A fuzzy clustering model for multivariate spatial time series. J. Classification 27 (1), 54–88.
Disegna, M., D'Urso, P., Durante, F., 2017. Copula-based fuzzy clustering of spatial time series. Spat. Stat. 21, 209–225.
D'Urso, P., De Giovanni, L., 2014. Robust clustering of imprecise data. Chemometr. Intell. Lab. Syst. 136, 58–80.
D'Urso, P., De Giovanni, L., Disegna, M., Massari, R., 2019a. Fuzzy clustering with spatial–temporal information. Spat. Stat. 30, 71–102.

D'Urso, P., Manca, G., Waters, N., Girone, S., 2019b. Visualizing regional clusters of Sardinia's EU supported agriculture: A Spatial Fuzzy Partitioning Around Medoids. Land Use Policy 83, 571–580.
D'Urso, P., Massari, R., Santoro, A., 2011. Robust fuzzy regression analysis. Inform. Sci. 181 (19), 4154–4174.
Emerson, J.D., Strenio, J., 1983. Boxplots and batch comparison. Underst. Robust Explor. Data Anal. 58.
Ester, M., Kriegel, H.-P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, KDD-96, Portland, Oregon, USA. pp. 226–231.
Everitt, B.S., Landau, S., Leese, M., Stahl, D., 2011. Cluster Analysis, fifth ed. Wiley, United Kingdom.
Fouedjio, F., 2016. A hierarchical clustering method for multivariate geostatistical data. Spat. Stat. 18, 333–351. http://dx.doi.org/10.1016/j.spasta.2016.07.003.
Fouedjio, F., 2017. A spectral clustering approach for multivariate geostatistical data. Int. J. Data Sci. Anal. 4 (4), 301–312.
García-Escudero, L.A., Gordaliza, A., Matrán, C., 2003. Trimming tools in exploratory data analysis. J. Comput. Graph. Statist. 12 (2), 434–449.
García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., 2008. A general trimming approach to robust cluster analysis. Ann. Statist. 36 (3), 1324–1345.
García-Escudero, L.A., Gordaliza, A., Matrán, C., Mayo-Iscar, A., 2010. A review of robust clustering methods. Adv. Data Anal. Classif. 4 (2–3), 89–109.
Gneiting, T., Kleiber, W., Schlather, M., 2010. Matérn cross-covariance functions for multivariate random fields. J. Amer. Statist. Assoc. 105 (491), 1167–1177.
Goovaerts, P., 1997. Geostatistics for Natural Resources Evaluation. Oxford University Press on Demand.
Gordon, A.D., 1999. Classification. Chapman and Hall/CRC, New York, http://dx.doi.org/10.1201/9780367805302.
Haslett, J., Bradley, R., Craig, P., Unwin, A., Wills, G., 1991. Dynamic graphics for exploring spatial data with application to locating global and local anomalies. Amer. Statist. 45 (3), 234–242.
Liu, H., Jezek, K.C., O'Kelly, M.E., 2001. Detecting outliers in irregularly distributed spatial data sets by locally adaptive and robust statistical analysis and GIS. Int. J. Geogr. Inf. Sci. 15 (8), 721–741.
Lu, C.T., Chen, D., Kou, Y., 2003. Detecting spatial outliers with multiple attributes. In: Proceedings. 15th IEEE International Conference on Tools with Artificial Intelligence. IEEE, pp. 122–128.
Oliver, M.A., Webster, R., 1989. A geostatistical basis for spatial weighting in multivariate classification. Math. Geology 21 (15), http://dx.doi.org/10.1007/BF00897238.
Pannatier, Y., 2012. VARIOWIN: Software for Spatial Data Analysis in 2D. Springer Science & Business Media.
Pebesma, E., 2004. Multivariable geostatistics in S: the gstat package. Comput. Geosci. 30, 683–691.
Pebesma, E.J., Bivand, R.S., 2005. Classes and methods for spatial data in R. R News 5 (2), 9–13, URL https://CRAN.R-project.org/doc/Rnews/.
Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Statist. Assoc. 66 (336), 846–850.
Romary, T., Ors, F., Rivoirard, J., Deraisme, J., 2015. Unsupervised classification of multivariate geostatistical data: Two algorithms. Comput. Geosci. 85, 96–103. http://dx.doi.org/10.1016/j.cageo.2015.05.019.
Rossiter, D.G., Zeng, R., Zhang, G.-L., 2017. Accounting for taxonomic distance in accuracy assessment of soil class predictions. Geoderma 292, 118–127.
Rousseeuw, P.J., 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65.
Schlather, M., Malinowski, A., Menck, P.J., Oesting, M., Strokorb, K., 2015. Analysis, simulation and prediction of multivariate random fields with package RandomFields. J. Stat. Softw. 63 (8), 1–25, URL http://www.jstatsoft.org/v63/i08/.
Shekhar, S., Chawla, S., 2003. A Tour of Spatial Databases. Prentice Hall Upper Saddle River.
Shekhar, S., Lu, C.T., Zhang, P., 2001. Detecting graph-based spatial outliers: algorithms and applications (a summary of results). In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, pp. 371–376.
Tukey, J.W., 1977. Exploratory Data Analysis. Addison-Wesley.
Wand, M.P., Jones, M.C., 1995. Kernel Smoothing. Chapman and Hall/CRC.
Wu, K.L., Yang, M.S., 2002. Alternative c-means clustering algorithms. Pattern Recognit. 35 (10), 2267–2278.