Stat Papers (2014) 55:29–47
DOI 10.1007/s00362-013-0524-z

REGULAR ARTICLE

Identification of local multivariate outliers

Peter Filzmoser · Anne Ruiz-Gazen · Christine Thomas-Agnan

Received: 16 April 2012 / Revised: 11 February 2013 / Published online: 3 May 2013

Abstract The Mahalanobis distance between pairs of multivariate observations is used as a measure of similarity between the observations. The theoretical distribution is derived, and the result is used for judging on the degree of isolation of an observation. In case of spatially dependent data where spatial coordinates are available, different exploratory tools are introduced for studying the degree of isolation of an observation from a fraction of its neighbors, and thus to identify local multivariate outliers.

Keywords Robust statistics · Spatial dependence · Outliers · MCD estimator · Mahalanobis distance

1 Introduction

Multivariate outlier detection belongs to the most important tasks for the statistical analysis of multivariate data. Their presence allows to draw conclusions about the data quality and about atypical phenomena in the data. Multivariate outliers behave differently than the majority of observations which are assumed to follow some underlying model like a multivariate normal distribution. The deviations of outlying observations from the majority of data points can also be understood in an exploratory context, e.g.

P. Filzmoser (✉)
Department of Statistics and Probability Theory, Vienna University of Technology, Vienna, Austria
e-mail: P.Filzmoser@tuwien.ac.at

A. Ruiz-Gazen · C. Thomas-Agnan
Toulouse School of Economics, Toulouse, France
e-mail: anne.ruiz-gazen@tse-fr.eu

C. Thomas-Agnan
e-mail: christine.thomas@tse-fr.eu

by visualizing a measure describing outlyingness and inspecting possible deviations or gaps in the resulting plot. Examples for such an approach are Atkinson and Mulira (1993), Atkinson et al. (2004), or Rousseeuw and Leroy (2003, Chap. 6).

The most commonly used measure of outlyingness is the Mahalanobis distance (Mahalanobis 1936). This multivariate distance measure assigns each observation a distance to the center, taking into account the multivariate covariance structure. Thus for observations $Z_1, \ldots, Z_n$ in the p-dimensional space with center μ and covariance Σ, the Mahalanobis distance is defined as

$$\mathrm{MD}_{\mu,\Sigma}(Z_i) = \left[(Z_i - \mu)^t \Sigma^{-1} (Z_i - \mu)\right]^{1/2} \quad \text{for } i = 1, \ldots, n. \tag{1}$$

Practically, for obtaining a reliable distance measure for multivariate data it is crucial how center μ and covariance Σ are estimated from the data. Classical estimates (arithmetic mean and sample covariance matrix) can be influenced by outlying observations, and thus robust estimates have to be used instead (Rousseeuw and Van Zomeren 1990; Maronna et al. 2006). A frequently used robust estimator of multivariate location and scatter is the Minimum Covariance Determinant (MCD) estimator. The MCD estimator looks for a subset of observations with smallest determinant of the sample covariance matrix. Rousseeuw and Van Driessen (1999) introduced a fast algorithm for computing the MCD estimator. As a cut-off value for the robust Mahalanobis distance the value $\sqrt{\chi^2_{p;0.975}}$ is suggested, which is the square root of the 97.5 % quantile of the chi-square distribution with p degrees of freedom. In the following we will use the notation $\chi_{p;0.975}$. Thus, values of the Mahalanobis distance larger than this cut-off value are considered as potential multivariate outliers. Note that there are also other proposals in the literature for finding an appropriate cut-off value, like an adaptive cut-off value that also takes into account sample size and dimension of the data, see Filzmoser et al. (2005).

The distance measure (1) for multivariate outlier detection does not account for any spatial dependence among the observations. Moreover, it is limited to identify overall, "global" outliers that differ from the main bulk of the data, but not necessarily outliers in a local neighborhood. The detection of such "local" spatial outliers is of interest in many fields where the data points have a spatial component, like in image analysis, in market segmentation, or in the statistical analysis of environmental data. Several approaches to local outlier identification for spatial data have been developed in computer vision and computer science (Haslett et al. 1991; Breunig et al. 2000; Chiu and Fu 2003; Papadimitriou et al. 2003). One of the goals of these exploratory tools is to detect "spatial outliers", i.e., observations that differ from their neighbors (Haslett et al. 1991; Cressie 1993, p. 33).

Interestingly, spatial or "local" outliers are most often also outlying according to the spatial dependence. Usually, it turns out that spatial data sets contain positive spatial autocorrelation which means that observations with high (respectively low) values for an attribute are surrounded by neighbors which are also associated with high (respectively low) values. Thus, in a positive autocorrelation scheme, observations that differ from their neighbors do not follow the same process of spatial dependence as the main bulk of the data. Graphics such as the variogram cloud (Cressie 1993) and

the Moran scatterplot (Anselin 1995) are interesting tools for detecting local outliers in a univariate framework. Cerioli et al. (1999) have used the forward search approach to identify spatial outliers in the univariate context, that is, extreme observations with respect to their neighboring values. However, up to our knowledge very few proposals have been made in the multivariate context.

The main objective of the present paper is to introduce new exploratory tools in order to detect outliers in multivariate spatial data sets. Our purpose is also to illustrate that if global outliers are present in the data set, they are usually also local outliers and they can completely mask other local outliers. The exploratory tools we introduce do not only detect both kinds of outliers but also give an insight into their global or/and local nature.

In Sect. 2 we introduce exploratory tools that compare pairwise distances in the variable or attributes space (i.e. pairwise Mahalanobis distances) and in the coordinate or geographic space (i.e. pairwise Euclidean distances). It turns out that this comparison can be interpreted as multivariate counterpart to the variogram cloud (Cressie 1993). The relation between global and pairwise Mahalanobis distances is formalized in Sect. 3, and distributional properties are derived. An exploratory tool for local outlier identification is introduced in Sect. 4 and applied in Sect. 5 to real data. The final Sect. 6 provides conclusions and outlook for further research in this area.

2 Generalization of univariate tools for local outlier detection

2.1 Illustrative example

Throughout the paper we will illustrate the proposed concepts with a small artificial data example shown in Fig. 1. We simulated $n = 100$ observations with two geographical coordinates in a square and two quantitative attributes. The left plot in Fig. 1 shows the two-dimensional data where the majority of the points come from a bivariate normal distribution. The ellipse corresponds to values of $\chi_{2;0.975} = 2.72$ of the robust Mahalanobis distance based on MCD location and scatter estimates. Hence, all squares and the filled rhomb are outside the ellipse and thus they are identified as global outliers. Figure 1 (right) shows the spatial X- and Y-coordinates of the data. For four selected points (shown by the filled symbols), circles are drawn that correspond to a Euclidean distance of 2 units from the points. All points within this distance are drawn with the corresponding open symbols and they can be considered as neighbors to the points in the center of the circles. Since the same symbols were used in the left plot of Fig. 1, it is possible to see the relation of the points in the variable space and in the coordinate space. The filled square and all its neighbors (at a distance of 2 units) are multivariate outliers. The filled rhomb is a multivariate outlier but not the neighbors. The filled triangle is on the boundary of the cut-off value 2.72, and the neighbors (open triangles) are far away in the variable space. Finally, the filled circle is in the center of the data cloud but its neighbors are very different in the variable space. Filled triangle and circle should thus be identified as local outliers because they neighboring points are very different. The filled rhomb and the filled square are already identified as global outliers and their neighbors are different for the rhomb but similar for the square.

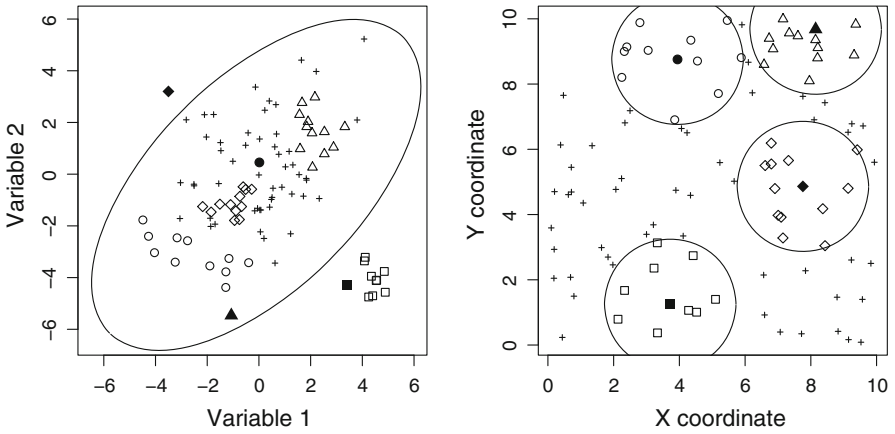

Fig. 1 Illustrative two-dimensional example with two spatial coordinates. *Left* Plot of the two variables with ellipse indicating global outliers; *right* plot of the spatial coordinates with *circles* indicating the neighborhood structure. The *same symbols* are used in both plots for the observations. The *filled rhomb* and the *filled square* are global outliers, the *filled circle*, the *filled triangle* and the *filled rhomb* are local outliers

Overall, we can distinguish the following cases:

(a) local, but not global outliers (filled triangle and filled circle);
(b) global, but not local outliers (filled square);
(c) local and global outliers (filled rhomb);
(d) neither local nor global outliers.

A reliable method for the identification of global and local outliers should be able to distinguish among these different situations.

2.2 Review of related approaches

Detecting and locating global and local outliers is one of the main objective of exploratory spatial data analysis. Graphics such as the Moran scatterplot (Anselin 1995) in econometrics and the variogram cloud (Chauvet 1982) in geostatistics are helpful in order to detect spatial outliers in a univariate context.

A Moran scatterplot rests on the definition of a neighborhood matrix which specifies in some sense the neighbors of each observation. In the econometrics literature, this neighborhood matrix is usually row-standardized so that the Moran scatterplot plots the values of the standardized observations in the abscissa and the values of the mean of the neighbors in the ordinate. Interpreting this scatterplot consists in comparing the value of an observation with the mean of its neighbors. When positive autocorrelation is present in the data set, the point cloud may exhibit some linear trend and the quadrants corresponding to positive/positive and negative/negative values contains the majority of the observations. So, local outliers may be found on the other quadrants, namely the positive/negative and negative/positive ones. But outliers may be also observations that exert high influence or leverage on the linear regression slope fitting the point cloud (Anselin 1996).

Anselin et al. (2002) proposed a multivariate version of the Moran scatterplot which consists in a scatterplot matrix with dimension the number of variables (2×2 in the bivariate case). Each scatterplot takes into account a couple of standardized variables (Z^k, Z^l) and plots the mean of the neighbors for Z^k on the ordinate versus the variable Z^l on the abscissa. Such a scatterplot matrix only takes into account bivariate relationships and its interpretation becomes intractable as soon as the number of variables is moderate to large. In that case, a multivariate alternative could be to calculate the Mahalanobis distance of each observation to the center of its neighbors as proposed in Lu et al. (2004). But this proposal does not lead to any graphical exploratory tool which could be interpreted as the Moran scatterplot. Another drawback of this approach appears as soon as global outliers are present in the data set. Because global outliers are likely to be associated with large Mahalanobis distances, they can completely mask other local outliers.

Another well-known exploratory tool, motivated by geostatistical ideas, is the variogram cloud (Cressie 1993; Haslett et al. 1991). For a pair (c_i, c_j) of data locations, $i, j = 1, \ldots, n, i \neq j$, let us consider the geographical Euclidean distance

$$\mathrm{ED}(c_i, c_j) = \left[(c_i - c_j)^t (c_i - c_j)\right]^{1/2}. \tag{2}$$

If the coordinates are univariate, this formula simplifies to $\mathrm{ED}(c_i, c_j) = |c_i - c_j|$.

The variogram cloud consists in plotting for all pairs (c_i, c_j) and for a single variable Z, the values $1/2\left(Z(c_i) - Z(c_j)\right)^2$ versus $\mathrm{ED}(c_i, c_j)$. The Euclidean distances may be calculated in some particular direction leading to a directional variogram cloud or in any direction leading to the so-called omnidirectional variogram cloud. Cressie (1993), claims that it is difficult to distinguish atypical observations from skewness using the variogram cloud and proposes to use the square-root differences cloud by replacing $\left(Z(c_i) - Z(c_j)\right)^2$ with $|Z(c_i) - Z(c_j)|^{1/2}$ on the ordinate axis.

In the multivariate setting, the most adequate tool for comparing observations is the Mahalanobis distance. So, our first proposal consists in a generalization of the variogram cloud for multivariate data by replacing the absolute differences with pairwise Mahalanobis distances, defined as

$$\mathrm{MD}_\Sigma(Z_i, Z_j) = \left[(Z_i - Z_j)^t \Sigma^{-1} (Z_i - Z_j)\right]^{1/2} \quad \text{for } i, j = 1, \ldots, n. \tag{3}$$

Similar to the "global" Mahalanobis distance (1), this distance measure between all pairs of observations accounts for the overall covariance structure.

The multivariate variogram cloud is a scatterplot of the $\mathrm{MD}_\Sigma(Z(c_i), Z(c_j))$ versus the geographical Euclidean distances $\mathrm{ED}(c_i, c_j)$ for $i, j = 1, \ldots, n$. Here we use the same indices for the observations in the variable space and in the coordinate space, i.e. $Z(c_i) = Z_i$ for $i = 1, \ldots, n$.

Using the illustrative example from above, the generalized variogram cloud for multivariate data is shown in Fig. 2 (left). The horizontal axis shows the pairwise Euclidean distances (2) and the vertical axis shows all pairwise Mahalanobis distances (3) using the MCD estimator for robustly estimating Σ. Since the illustrative data set consists of $n = 100$ observations, the plot shows $n * (n - 1)/2 = 4{,}950$ points

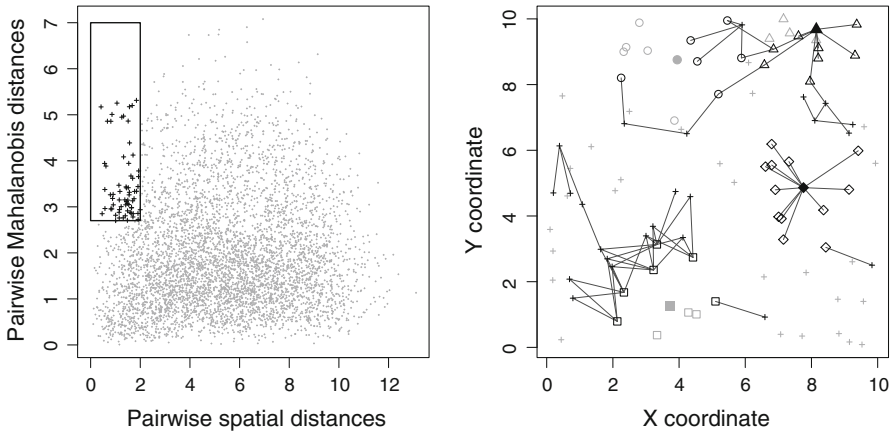

Fig. 2 Multivariate variogram cloud (*left*) for the illustrative example shown in Fig. 1. Potential local outliers are marked in the *left plot*, resulting in the linked pairs of observations in the coordinate plot (*right*)

representing all pairs (Z_i, Z_j) for $i < j$. A region in the plot is selected, and the points falling into this region are linked by lines in the plot of the coordinates (Fig. 2, right). The marked region selects observations with low pairwise spatial and large pairwise Mahalanobis distances, i.e. it should highlight potential local outliers. Figure 2 (right) does not highlight all local outliers and their closest neighbors. For the triangles this works quite well, but not for the circles. Moreover, the rhombs are clearly connected, but the filled rhomb is a global outlier. Finally, the filled square as global outlier is not highlighted, but only some of its neighbors. The picture would certainly be different by marking a larger region in the left plot. However, these results already indicate that the multivariate extension of the variogram cloud does not allow to distinguish between local and global outliers.

In the above procedure we did not differentiate the potential local outliers with respect to their location in the data cloud. For example, a local outlier in the center of the data cloud (like the filled circle in Fig. 1, left) may be differently treated than a local outlier on the boundary (the filled triangle). This is because in the center we expect a higher point density than on the boundary, and thus for local outlyingness central points may be more similar to their neighbors than boundary points. A more formalized approach of this statement will be presented in the following section.

3 Theoretical properties of the pairwise Mahalanobis distances

Let us consider a sample $Z_1, \ldots, Z_n$ of i.i.d. random vectors in p dimensions following a Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$ with $\mu \in R^p$ and Σ a $p \times p$ symmetric positive definite matrix. Hereafter, we denote by Z_i^k, $k = 1, \ldots, p$, the components of the random vector Z_i. The transformed random vectors $Y_i = \Sigma^{-1/2}(Z_i - \mu), i = 1, \ldots, n$, are i.i.d. and follow a Gaussian distribution $\mathcal{N}_p(0, I)$ where I denotes the $p \times p$ identity matrix.

According to (1) the squared Mahalanobis distance between $Z_i, i = 1, \ldots, n$, and the location parameter μ is:

$$\mathrm{MD}^2_{\mu,\Sigma}(Z_i) = (Z_i - \mu)^t \Sigma^{-1}(Z_i - \mu) = Y_i^t Y_i.$$

It is well known that the $\mathrm{MD}^2_{\mu,\Sigma}(Z_i)$, $i = 1, \ldots, n$, are i.i.d. and follow a χ^2_p distribution.

Due to (3) the pairwise squared Mahalanobis distance between Z_i, and Z_j, $i, j = 1, \ldots, n$ is:

$$\begin{aligned}\mathrm{MD}^2_{\Sigma}(Z_i, Z_j) &= (Z_i - Z_j)^t \Sigma^{-1}(Z_i - Z_j) \\ &= (Y_i - Y_j)^t (Y_i - Y_j).\end{aligned}$$

Proposition 1 *If we consider i.i.d. Gaussian random vectors $Z_1, \ldots, Z_n$, the conditional distribution of the pairwise squared Mahalanobis distances $MD^2_{\Sigma}(Z_i, Z_j)$, $j = 1, \ldots, n$, given Z_i is a non-central chi-square distribution with p degrees of freedom and the non-centrality parameter $MD^2_{\mu,\Sigma}(Z_i)$.*

Proof We consider the conditional distribution of $\mathrm{MD}^2_{\Sigma}(Z_i, Z_j)$ when $Z_i = z$ with $z = (z^1, \ldots, z^p)^t \in R^p$. Let $y = \Sigma^{-1/2}(z - \mu) = (y^1, \ldots, y^p)^t$. We have

$$\begin{aligned}\mathrm{MD}^2_{\Sigma}(Z_j, z) &= (Z_j - z)^t \Sigma^{-1}(Z_j - z) \\ &= (Y_j - y)^t (Y_j - y) \\ &= \sum_{k=1}^{p} \left(Y_j^k - y^k\right)^2.\end{aligned}$$

We know that if Z^k, $k = 1, \ldots, p$, are p independent normally distributed random variables with mean μ_k and variance σ_k^2, then the random variable $\sum_{k=1}^{p} \left(Z^k/\sigma_k\right)^2$ is distributed according to the non-central chi-square distribution with p degrees of freedom and non-centrality parameter $\lambda = \sum_{k=1}^{p} (\mu_k/\sigma_k)^2$ (see, e.g., Evans et al. 1993, p. 51). We will use the notation $\chi^2_p(\lambda)$ for this distribution.

We have the terms $(Y_j^k - y^k)$, $k = 1, \ldots, p$, which are independent normally distributed with mean $-y^k$ and variance 1. So, the $\mathrm{MD}^2_{\Sigma}(Z_j, z)$, $j = 1, \ldots, n$ follow a non-central chi-square distribution with p degrees of freedom and non-centrality parameter

$$\lambda = \sum_{k=1}^{p} (-y^k)^2 = y^t y = (z - \mu)^t \Sigma^{-1}(z - \mu) = \mathrm{MD}^2_{\mu,\Sigma}(z),$$

or, in our notation, $\chi^2_p(\mathrm{MD}^2_{\mu,\Sigma}(z))$. □

Using the result of Proposition 1 it is possible to define outlyingness in a local sense. In the following we will use the notation $z_1, \ldots, z_n$ for the sample values, $\mathrm{MD}(z_i)$ for the Mahalanobis distances ($i = 1, \ldots, n$), and $\mathrm{MD}(z_i, z_j)$ for the pairwise Mahalanobis distances ($i, j = 1, \ldots, n$). The Mahalanobis distances and the pairwise Mahalanobis distances are estimated robustly using the MCD estimator.

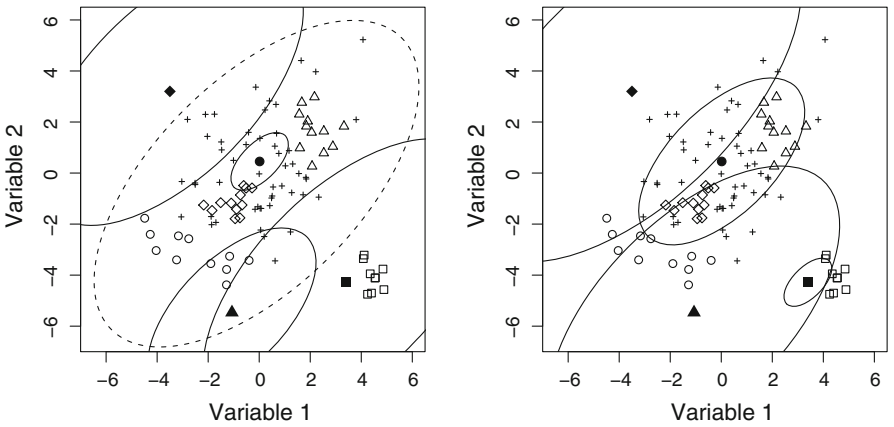

Fig. 3 Illustrative example from Fig. 1. *Left* Cut-off value for global outliers (*dashed ellipse*) and 10 % quantiles of the non-central chi-square distribution for the points with the *filled symbols* (*ellipses with solid lines*). *Right* Quantiles of the non-central chi-square distribution adjusted to include the next neighbor

Since local outliers are supposed to be different from their neighbors, one could define a quantile of the non-central chi-square distribution and count the number of neighbors falling into this defined range. This concept is visualized in Fig. 3 (left) for the example data presented in Fig. 1. The dashed ellipse visualizes the cut-off value $\chi_{2;0.975}$ for global outliers, see Fig. 1 (left). The other four ellipses correspond to the values $\chi_{2;0.1}(\mathrm{MD}^2(z_i))$ where z_i, $i = 1, \ldots, 4$ are the four points with the filled symbols in the centers of these ellipses. Here the 10 % quantile was chosen for inspecting the neighborhood, but any other value could also be selected. In case of independence and normal distribution we would expect 10 % of the values falling inside an ellipse. Consequently, the ellipses in the center of the data cloud are smaller than on the boundary which is according to the non-centrality parameter of the chi-square distribution. Figure 3 (left) shows that none of the neighbors (defined by the circles in Fig. 1, right) falls into the ellipses, except for the filled square where all neighbors (open squares) are inside the 10 % quantile. Thus, the filled circle and the filled triangle are local outliers because they are isolated from their neighbors, the filled rhomb is an isolated global outlier, and the filled square is a clustered global outlier.

Remark The assumption of independence used in Proposition 1 will not be valid in particular for spatially dependent data. Dale and Fortin (2009) discuss this issue in the context of statistical testing for data with spatial autocorrelation. Moreover, the assumption that Σ is the same for all observations is a simplification that would imply some form of stationarity, in particular if the i.i.d. assumption is violated. Furthermore, the distribution derived in Proposition 1 is valid either if Σ is known, or if n diverges. In practice, however, we deal with a limited number of observations, and we have to estimate Σ. Here, Σ will be estimated robustly, using the MCD estimator. The effect of estimating the covariance matrix with the MCD estimator on the distribution of the resulting Mahalanobis distances was addressed in several papers, like in Hardin

and Rocke (2005), Riani et al. (2009), Cerioli (2010), or Cerioli et al. (2012). These authors derived better approximations of the distribution in order to obtain a more accurate cut-off value for outliers. However, since here we deal with the additional problem of violation from the i.i.d. assumption, we stick to the distribution and cut-off value proposed above, and note that this procedure for identifying local multivariate outliers must be understood in an exploratory context rather than as a precise statistical test. As an empirical evidence of this approach, it can be seen e.g. from Fig. 3 that our approach fulfills the purpose.

The above characterization of outlyingness depends on the considered quantile of the non-central chi-square distribution. However, local outlyingness could also be defined differently, by measuring the distance to the next neighbor in terms of the non-central chi-square distribution. This is illustrated in Fig. 3 (right). We compute for the four points with filled symbols the sizes of the ellipses needed to "touch" the next neighbor (with the corresponding open symbol). Let z_j be the next neighbor of z_i, i.e. the distance of c_j and c_i is the smallest among all the neighbors of observation z_i. According to Proposition 1, the pairwise squared Mahalanobis distance $\mathrm{MD}^2(z_i, z_j)$ is equal to a certain $\alpha(j)$-quantile $\chi^2_{2;\alpha(j)}(\mathrm{MD}^2(z_i))$ of the non-central chi-square distribution. The probabilities $\alpha(j)$ can then be easily determined, and the results for our example are: 59 % (filled circle), 43 % (filled triangle), 26 % (filled rhomb), and 0.02 % (filled square). These values give a good impression about the degree of isolation from the next neighbor.

Since just by chance the next neighbor could be close but a third neighbor far away, it can be more sensible to search for α-quantiles such that the corresponding ellipses include a pre-defined percentage, e.g. 10 %, of the next neighbors.

4 Tools for identifying local outliers

4.1 What are local outliers?

A characterization of local outliers requires a definition of the local neighborhood. For this purpose, two concepts are common, namely to fix a maximum distance $d_{\max}$ in the space of the spatial coordinates, and to define the neighbors of an observation z_i as all points z_j ($j = 1, \ldots, n$; $j \neq i$) where the distance $d_{i,j}$ between z_i and z_j is not larger than $d_{\max}$. As distance measure $d_{i,j}$ the Euclidean distance can be considered, see (2). A second concept is to define neighborhood by the nearest k observations. For finding the k nearest neighbors (kNN) of an observation z_i we have to consider the sorted distances $d_{i,(1)} \leq d_{i,(2)} \leq \cdots \leq d_{i,(k)} \leq d_{i,(n)}$ to all other observations. The kNN to z_i are all observations where $d_{i,j} \leq d_{i,(k)}$ for $j = 1, \ldots, n$, $j \neq i$.

Using a neighborhood based on $d_{\max}$ sometimes results in difficulties at the boundary of an area where usually less neighbors to an observation are found than away from the boundary. This is avoided for kNN where the number of neighbors is always fixed with k, regardless of the location of an observation.

The neighbors of an observation z_i are all observations z_j with $j \in N_i = \{i_1, \ldots, i_{n(i)}\}$. Clearly, if kNN is used to define the neighbors, the number of neighbors $n(i) = k$ for $i = 1, \ldots, n$. Local outlyingness of an observation implies that the

observation is very different from most of its neighbors. Therefore, β will denote a fraction, and $\lceil n(i) \cdot \beta \rceil$ is the number of neighbors of z_i that can be similar to z_i but the remaining neighbors have to be reasonably different (here, $\lceil x \rceil$ means rounding to an integer not smaller than x). Note that $0 \leq \beta < 0.5$ aims at looking for local outliers, but for $0.5 \leq \beta \leq 1$ it is possible to search for homogeneous regions.

Let $\mathrm{MD}^2(z_i, z_{(j)})$ denote the sorted squared pairwise Mahalanobis distances of observation z_i to all neighbors z_j, with $j \in N_i = \{i_1, \ldots, i_{n(i)}\}$. Similar to the previous section, the degree of isolation of an observation z_i from a fraction $(1 - \beta)$ of its neighbors can be characterized by the $\alpha(i)$-quantile

$$\chi^2_{p;\alpha(i)}(\mathrm{MD}^2(z_i)) = \mathrm{MD}^2(z_i, z_{(\lceil n(i)\cdot\beta \rceil)}) \quad \text{for } i = 1, \ldots, n. \tag{4}$$

$\alpha(i)$ measures the local outlyingness of an observation z_i. For a large number of neighbors, and in case of independence and normal distribution, $\alpha(i)$ should approximate β. However, if $\alpha(i)$ is substantially larger than β, observation z_i is considered as potential local outlier.

This characterization of local outliers depends on the size of the neighborhood ($d_{\max}$ or k), and on the fraction β. For the exploratory tools for local outlier identification introduced in the following we either need to fix the fraction β of neighbors, or the maximum distance $d_{\max}$ (alternatively k for kNN) in order to define the neighborhood size, or both. The plots in the following demonstrate the ideas using the illustrative example of Fig. 1. We use kNN for defining the neighborhood since the concept of maximum distance could give instable results for small values of $d_{\max}$ because of the small sample size. In all examples we use the MCD estimator to estimate the covariance matrix Σ.

4.2 Variable neighborhood size and fixed fraction β of neighbors

In the first tool for local outlier identification we vary the number of neighbors of each observation using kNN with $k = 1, \ldots, 99$. So, if $k = 99$, all observations are neighbors of any observation except itself. The fraction β is fixed with 10 %. For each observation we are computing the degree of isolation from $1 - \beta = 90\,\%$ of its neighbors, using Eq. (4). Figure 4 shows the results in two separate plots for the regular observations and for (global) outlying observations. This separation avoids the confusion of local outliers that are not global ones, and global outliers that could also be local ones. Each line in the plots belongs to one specific observation. In both plots we marked two lines (black) which are the two constructed local outliers in the left plot (filled circle and triangle in Fig. 1) and the two constructed global outliers in the right plot (filled square and rhomb in Fig. 1).

All outlier detection tools proposed in this paper are implemented in the R package *mvoutlier* (Filzmoser and Gschwandtner 2012). The data of the toy example are available in the package via `data(dat)` (bivariate data), and `data(X)`, `data(Y)` (spatial coordinates). Figure 4 can be generated by:

```
locoutNeighbor(dat,X,Y,propneighb=0.1,chisqqu=0.975,
variant=''knn'',usemax=1,npoints=100,indices=c(1,11,24,36))
```

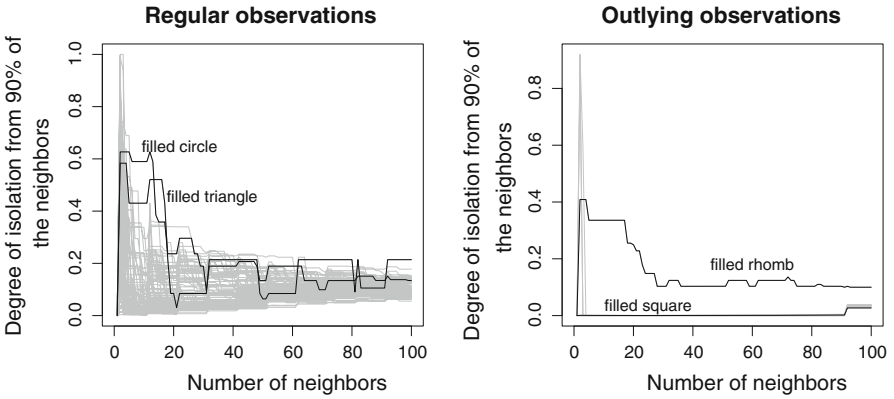

Fig. 4 Degree of isolation of each observation (*lines*) from 90 % of the neighbors. The size of the neighborhood is changed (*horizontal axes*). Separate plots are drawn for regular (*left*) and outlying (*right*) observation. The data are from the illustrative example shown in Fig. 1

The parameter `propneighb` corresponds to the fraction β, `chisqqu` provides the quantile for the chi-square distribution separating regular observations from outliers, `variant` defines the distance measure used (here kNN), `usemax` controls the fraction of closest points that are used in the plot (here all neighbors are used), `npoints` defines the number of points for which the calculations are done (here at each of the 100 data points), and `indices` highlights the observations with the corresponding indices in the plot (these are the indices of the observations with the filled symbols).

For very small values of k we observe some instability. The reason is that just by chance two observations could be close in the spatial sense but very different in the variable space. For a larger neighborhood the local outlier measure becomes more reliable. The two black lines in the left plot of Fig. 4 are exceptionally high for small neighborhoods, i.e. their degree of isolation $\alpha(i)$ is substantially larger than $\beta = 10\,\%$. There are also other observations that have a high degree of isolation for certain neighborhood sizes. This can of course happen since we did not control the structure of the observations marked with "+" in Fig. 1. The two marked outliers in the right plot have different behavior; one is identified as local outlier with a high degree of isolation (filled rhomb) and the second is a clustered outlier (filled square).

This plot can give an idea at which size of the neighborhood local outliers are highlighted. For example, $k = 10$ neighbors flags the constructed outliers quite well.

4.3 Fixed neighborhood size and variable fraction β of neighbors

The plots in Fig. 4 have shown exceptional behavior of some observations for a neighborhood size of $k = 10$. Now we fix the value $k = 10$ but vary the fraction β to compute the degree of isolation. Figure 5 shows the resulting plots for the regular observations (left) and for the outliers (right). The horizontal axes represent β and the vertical axes are the degrees of isolation, see (4). We marked the same observations as in Fig. 4 with dark lines. Obviously, the results in Fig. 4 for $k = 10$ neighbors

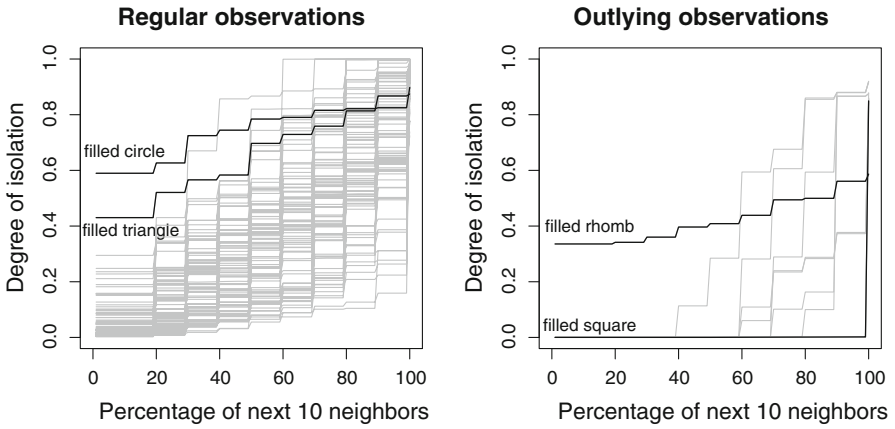

Fig. 5 Degree of isolation of each observation (*lines*) from a varying percentage β of the next 10 neighbors. Separate plots are drawn for regular (*left*) and outlying (*right*) observation. The data are from the illustrative example shown in Fig. 1

which were computed for $\beta = 10\,\%$ correspond to the results for $\beta = 10\,\%$ in Fig. 5. Here we get additional information about the isolatedness of each observation from a varying percentage of the nearest 10 neighbors. The two marked observations in the left plot are very isolated already from all 10 neighbors, but are still isolated from half of their next 10 neighbors. The right plot reveals again the filled rhomb as very isolated local outlier, and the filled square as outlier being very similar to the next 10 neighbors.

Figure 5 can be generated in R by:

```
locoutPercent(dat,X,Y,k=10,chisqqu=0.975,indices=c(1,11,24,36))
```

The parameter `k` controls the considered number of next k neighbors, and the other parameters are the same as explained in the R code to Fig. 4.

At each fixed value of β we obtain an order of the observations according to their degree of isolation. This ordering will be used in the next tool.

4.4 Fixed neighborhood size and fixed fraction β of neighbors

The plots in Figs. 4 and 5 allowed to pursue the degree of isolation of each observation by varying k and β. We fix these parameters for the plots in Fig. 6, and we select $k = 10$ and $\beta = 10\,\%$ because these values were revealing our constructed local outliers. The left plot in Fig. 6 shows the sorted index of the observations, where the sorting is from smallest (left) to highest (right) degree of isolation, against the pairwise Mahalanobis distance to the k neighbors. Thus each point in the plot represents one neighbor to a selected observation, and all k neighbors to a specific observations are arranged on the same horizontal position (Sorted index). The plot window is split by a vertical line into two parts, the left part for the regular observations, and the right plot for the outliers. In this plot we can interactively select regions. Here we selected

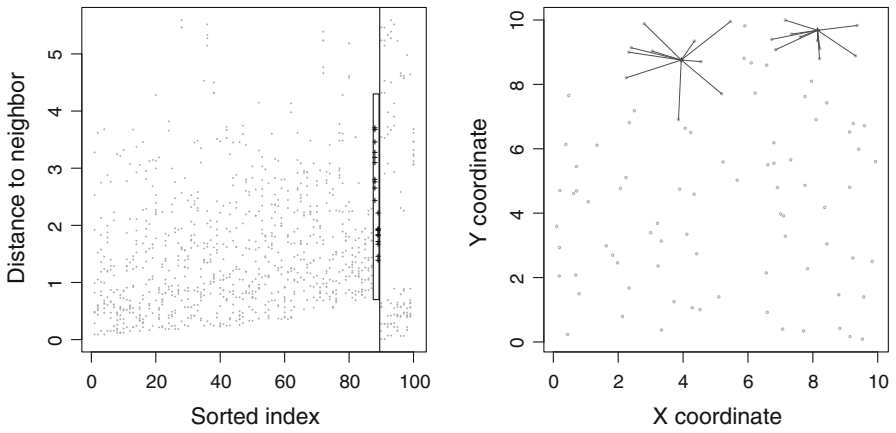

Fig. 6 *Left* Interactive plot of the observation index sorted by the degree of isolation versus the distance to all neighbors. *Right* Spatial coordinates of the observations. The marked points in the *left plot* are connected by *lines* in the *right plot*. The data are from the illustrative example shown in Fig. 1

for the two most isolated regular observations all neighbors. The right plot in Fig. 6 shows the spatial coordinates of the observations, and the selected observations and their neighbors are connected by lines. Thus it is easy to identify the observations with highest degree of isolation, the potential local outliers, together with their neighbors. These are the two constructed local outliers of Fig. 1 and the neighbors thereof.

Figure 6 can be generated in R by:

```
locoutSort(dat,X,Y,k=10,propneighb=0.1,chisqqu=0.975)
```

The parameters are the same as used earlier for Figs. 4 and 5. By defining an area interactively in the generated plot (Fig. 6, left), points can be selected and analyzed in the next plot (Fig. 6, right).

In Fig. 6 one could mark more observations in order to study the local behavior. One could also select observations from the global outliers to study their local outlyingness. Furthermore, it is possible to identify regions that are locally homogeneous. For this purpose the left plot of Fig. 6 has to be made for a larger value of β, e.g. for $\beta = 0.9$. Locally homogeneous observations are in the lower left part where the indices refer to observations which are very similar to a fraction β of the nearest k neighbors.

5 Application to real data

5.1 Geochemical data from northern Europe

The so-called Baltic Soil Survey (BSS) data originates from a large-scale geochemistry project carried out in northern Europe, in an area of about 1,800,000 km^2 (Reimann et al. 2003). In two different layers, 769 samples of agricultural soils have been collected. The samples were analyzed for the concentration of more than 40 chemical elements. This project was carried out to document element concentrations and spatial variation

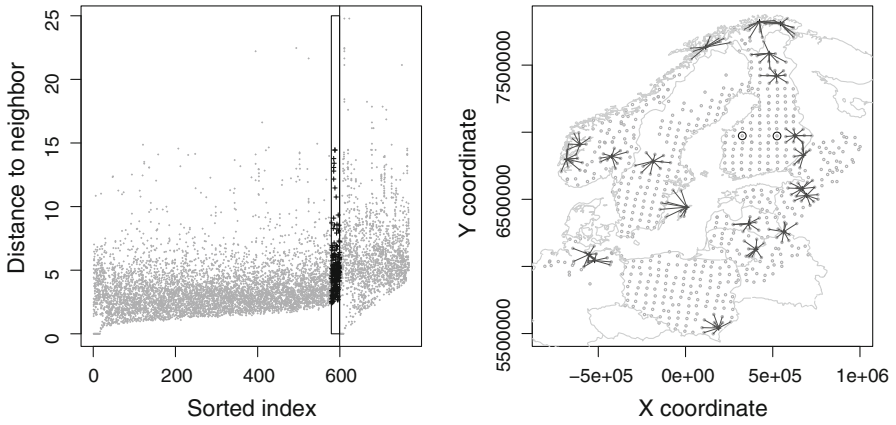

Fig. 7 Plots for identifying local outliers in the BSS data set. The 20 observations and their neighbors ($k = 10$) with the highest degree of isolation ($\beta = 10\,\%$) are highlighted

in agricultural soils from northern Europe. The element distributions will not only be influenced by the underlying lithology but also by other factors like climate, the input of marine aerosols, agricultural practice and contamination. The data sets of both layers are available in the R package *mvoutlier* as data files *bsstop* and *bssbot*.

As an example for local outlier identification we use the elements Al_2O_3, Fe_2O_3, K_2O, MgO, MnO, CaO, TiO_2, Na_2O, P_2O_5 and SiO_2 from the top layer (0–25 cm). These are major elements and their element concentrations sum up to almost 100 %. In the literature, this kind of data is known as compositional data (Aitchison 1986). Before continuing with outlier detection an appropriate transformation has to be made, and therefore we use the isometric logratio transformation (Filzmoser and Hron 2008; Filzmoser et al. 2012). We apply the interactive plot from Fig. 6 by selecting the parameters $k = 10$ and $\beta = 10\,\%$. So, we are interested in deviations in small neighborhoods, which, because of the large scale, still correspond to quite big areas. Highlighting the 20 observations and their neighbors with the highest degree of isolation (of the regular observations) results in the plots in Fig. 7.

The identified points in the map of Fig. 7 are mainly boundary points which have a higher probability of local outlyingness. Some of these observations are known to behave differently, e.g. those on the northern coast. However, finding the reason for local outlyingness of all marked points would require much more detailed study of the single element maps. Reasons for their abnormal behavior could be a different data structure caused by local artefacts in the soil, or exchanged samples, incorrect sample preparation, wrong laboratory analyses, etc. Rather than going into detail with studying the identified points, we want to test the proposed procedures by exchanging the spatial locations of two samples. This sometimes happens in the laboratory where the chemical analysis of the samples was correctly done, but the assignment of the samples to the locations was mixed up. The samples to be exchanged are marked by circles in the right plot of Fig. 7. Figure 8 shows the result where we again marked the 20 observations and their neighbors with the highest degree of isolation. Both of the exchanged samples appear now as local outliers, or more precisely, they are among the 20 most isolated regular observations. This means that their multivariate data structure

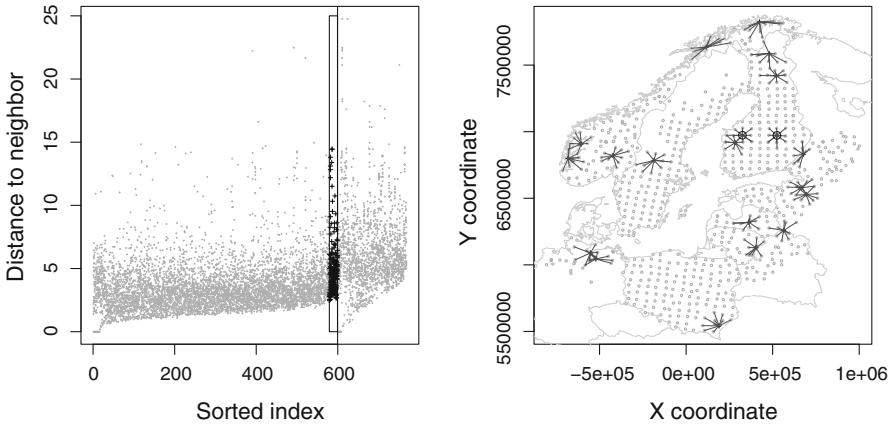

Fig. 8 Plots for identifying local outliers in the BSS data set. The 20 observations and their neighbors ($k = 10$) with the highest degree of isolation ($\beta = 10\,\%$) are highlighted. The observations marked by *circles* were exchanged

is now quite different from their neighbors, although the exchanged samples are not that distant (around 200 km) compared to the size of the whole survey area.

We are also interested in the actual degree of isolation in order to get an idea of the strength of local outlyingness. Since we selected $\beta = 10\,\%$ we would expect a degree of isolation of about $10\,\%$ in case of independent normally distributed data. Using the tools of Fig. 5 we inspect the behavior of the original and manipulated data. Since we were only investigating the regular observations, Fig. 9 shows the appropriate plots only for the regular observations. The left plot shows the original data where the observations that will be exchanged are marked by black lines. They are very homogeneous to their neighbors. In fact there are only about 10 observations being exceptionally different from their neighbors. The right plot shows the graph for the manipulated data, with the exchanged observations marked by black lines. They clearly increased the degree of isolation, even for a wide range of β, and therefore they are identified as local outliers.

In a further experiment with this data set, we test the local outlier identification procedure by means of a small simulation study. This time we randomly pick up two observations, which could be regular or outlying observations. For each observation we compute their rank of isolation. A rank of one refers to the most isolated observation (of the regular or outlying points). We store the minimal rank of both observations. Then the spatial coordinates of these observations are exchanged. Again, the minimal rank of the exchanged observations is computed. This procedure is repeated 100 times, and the results are shown in Fig. 10. On the horizontal axis, the distance between the two selected observations is shown. The points with the dark symbols refer to the minimal rank of the pairs for the original data. The lines connecting to the open symbols refer to the minimal rank, as it occurs after exchanging the coordinates. It can be seen that in almost all cases, the rank of isolatedness gets much lower after exchanging the coordinates. Most of the points fall below rank 20, indicated by the dashed horizontal line, which means that at least one of these points will very likely be identified as

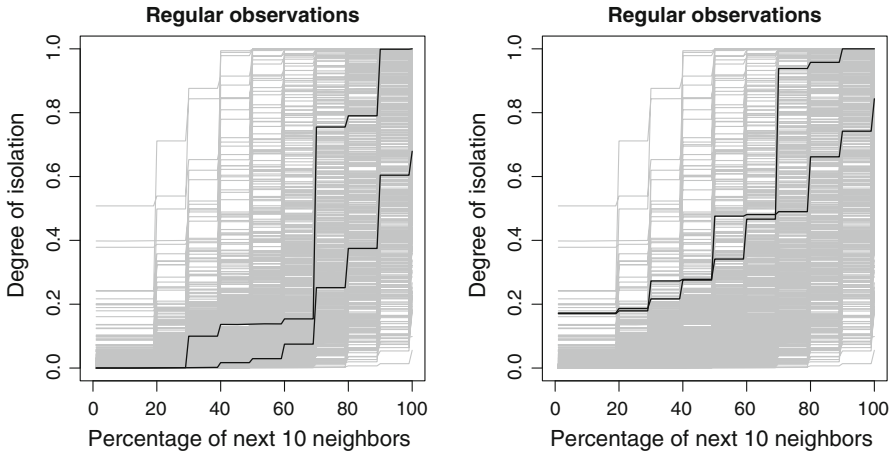

Fig. 9 Plots for visualizing the degree of isolation for each observation of the BSS data set. The *marked lines* refer to two observations for which the spatial coordinates have been exchanged in the *right plot*

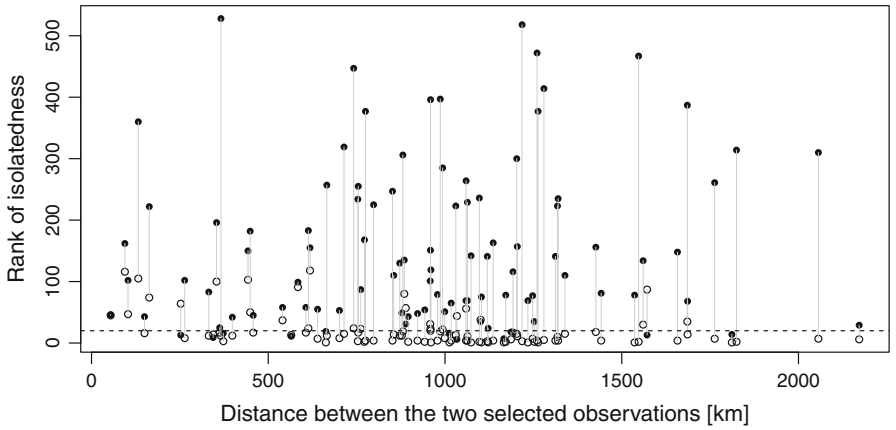

Fig. 10 Pairs of observations are selected randomly, and their isolatedness is ranked. The minimum of their rank is shown by *dark symbols*, as well as the spatial distance of the pairs. The connected points with the *open symbols* show the (minimum) rank after exchanging the spatial coordinates. The *horizontal line* indicates rank 20

local outlier. If the distance between both points is very low, the rank improvement is somewhat smaller.

5.2 Social data from France

The data set considered here originates from Guerry (1833), who collected and analyzed social data from the 86 departments of France around 1830, with the view to determining social laws. The data set is available in the R package *Guerry*. In the following we will use only 85 departments by excluding Corsica (which is an outlier

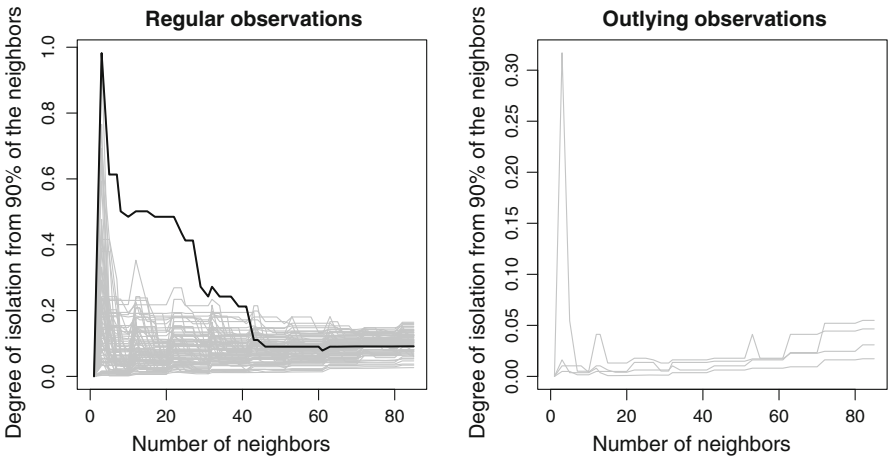

Fig. 11 Degree of isolation of each observation (*lines*) from 90 % of the neighbors for the social data

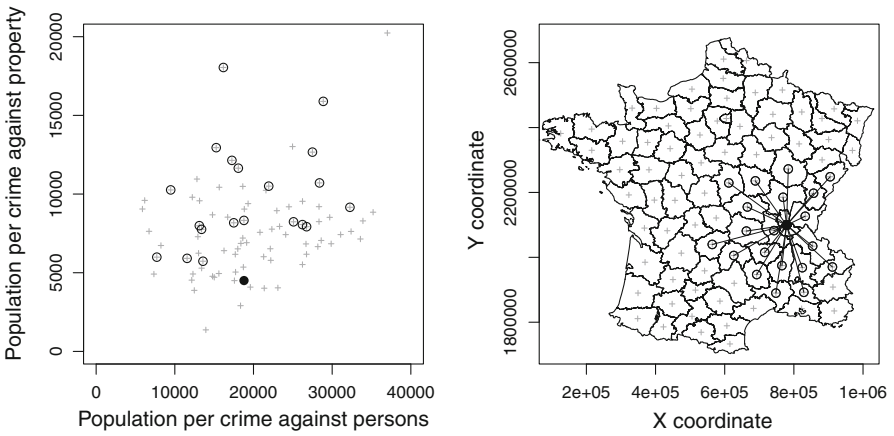

Fig. 12 Identification of the most extreme local outlier and its next 20 neighbors

both spatially and statistically), and the two variables "population per crime against persons" and "population per crime against property". Considering only two variables allows to graphically inspect the data, see Fig. 12 (left). Figure 12 (right) shows the 85 departments, where the indicated points are used as spatial coordinates.

Using the ideas from Sect. 4.2 we can analyze the degree of isolation from a fraction $1-\beta$ of the neighbors. We will consider again $\beta = 10\,\%$, and vary the neighborhood size with kNN. The resulting plots are shown in Fig. 11. In the right plot we can see that four global outliers are identified, which we will not further pursue here. The left plot for the regular observations reveals one observation as clearly deviating, even for a large range of next neighbors. We will analyze this local outlier in the following.

Selecting this most isolated local (but not global) outlier in the same way as in Figs. 6, 7 and 8, by considering the $k = 20$ next neighbors, gives the plot in Fig. 12 (right). The local outlier is shown by the filled circle, the 20 neighbors are linked and marked

by circles. In the left plot we show here the original data, where the same symbols are used as in the right plot. Indeed, the local outlier is very different from most of its neighbors. The region of the outlier and its neighbors corresponds more or less to the Rhône-Alpes region, an alpine region next to Switzerland. The local outlier is the department Rhône with the capital Lyon. In the nineteenth century, Lyon was already a famous industrial city, with focus on silk production. The region around Lyon was densely populated, quite different from most of its neighboring departments. This might be possible reasons why the crime rates are different to most other neighbors.

6 Conclusions

When analyzing multivariate data it is of interest to identify outliers or other relevant structures in the data set. Here we assume the additional availability of spatial coordinates (one-, two- or three-dimensional) for each multivariate observation. Spatial dependence is in fact a quite frequent situation, but often the coordinates are either not reported or ignored for the analysis. We have introduced different exploratory tools to identify outliers in a local spatial neighborhood. The tools are based on pairwise robust Mahalanobis distances between the observations. The robustness of the method is the result of plugging in a robust covariance estimation for computing the Mahalanobis distances. The determination of the distribution of these pairwise distances allows deriving a measure for local outlyingness. The local behavior of the method can be regulated by changing the size of the neighborhood. Additional "robustness" for identifying local outliers is included by tolerating a (small) percentage of similar neighbors, which can be similar just by chance. By increasing this percentage, the tools can even be used for finding locally homogeneous regions.

Acknowledgments The ideas are not limited to data with spatial coordinates; they could also be extended to time series data, or data with spatial and temporal dependence. These will be tasks for future research. This work was supported by the French Agence Nationale de la Recherche through the ModULand project (ANR-11-BSH1-005).

References

Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London

Anselin L (1995) Local indicators of spatial association. Geogr Anal 27(2):93–115

Anselin L (1996) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer M, Scholten H, Unwin D (eds) Spatial analytical perspectives on GIS. Taylor and Francis, London, pp 111–125

Anselin L, Syabri I, Smirnov O (2002) Visualizing multivariate spatial correlation with dynamically linked windows. In: Anselin L, Rey S (eds) New tools for spatial data analysis: proceedings of a workshop, Center for Spatially Integrated Social Science, University of California, Santa Barbara (CD-ROM)

Atkinson AC, Mulira H-M (1993) The stalactite plot for the detection of multivariate outliers. J Stat Comput 3(1):27–35

Atkinson AC, Riani M, Cerioli A (2004) Exploring multivariate data with the forward search. Springer, New York

Breunig MM, Kriegel H-P, Ng RT, Sander J (2000) LOF: identifying density-based local outliers. In: Proceedings of the ACM SIGMOD (2000) international conference on management of data, Dallas, TX, pp 93–104

Cerioli A (2010) Multivariate outlier detection with high-breakdown estimators. J Am Stat Assoc 105:147–156

Cerioli A, Riani M (1999) The ordering of spatial data and the detection of multiple outliers. J Comput Graph Stat 8:239–258
Cerioli A, Farcomeni A, Riani M (2012) Robust distances for outlier-free goodness-of-fit testing. Comput Stat Data Anal (in press)
Chauvet P (1982) The variogram cloud. In: Proceedings of the 17th APCIM symposium, Colorado Scholl of Mines, Golden, April 19–23, 1982, pp 757–764
Chiu AL, Fu AW (2003) Enhancements on local outlier detection. In: Proceedings of the seventh international database engineering and applications, symposium (IDEAS'03), pp 298–307
Cressie N (1993) Statistics for spatial data. Wiley, New York
Dale MRT, Fortin M-J (2009) Spatial autocorrelation and statistical tests: some solutions. J Agric Biol Environ Stat 14(2):188–206
Evans M, Hastings N, Peacock B (1993) Statistical distributions, 2nd edn. Wiley, New York
Filzmoser P, Gschwandtner M (2012) mvoutlier: multivariate outlier detection based on robust methods. R package version 1.9.8. http://CRAN.R-project.org/package=mvoutlier
Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Math Geosci 40(3):233–248
Filzmoser P, Garrett RG, Reimann C (2005) Multivariate outlier detection in exploration geochemistry. Comput Geosci 31:579–587
Filzmoser P, Hron K, Reimann C (2012) Interpretation of multivariate outliers for compositional data. Comput Geosci 39:77–85
Guerry A-M (1833) Essai sur la statistique morale de la France. Crochard, Paris. English translation: HP Whitt and VW Reinking, Edwin Mellen Press, Lewiston, 2002
Hardin J, Rocke DM (2005) The distribution of robust distances. J Comput Graph Stat 14:910–927
Haslett J, Bradley R, Craig P, Unwin A, Wills G (1991) Dynamic graphics for exploring spatial data with applications to locating global and local anomalies. Am Stat 45(3):234–242
Lu CT, Chen D, Kou Y (2004) Multivariate spatial outlier detection. Int J Artif Intell Tools 13(4):801–812
Mahalanobis PC (1936) On the generalised distance in statistics. In: Proceedings of the National Institute of Science of India A2, pp 49–55
Maronna R, Martin D, Yohai V (2006) Robust statistics: theory and methods. Wiley Canada Ltd, Toronto
Papadimitriou S, Kitawaga H, Gibbons PB, Faloutsos C (2003) LOCI: fast outlier detection using the local correlation integral. In: Dayal U, Ramamritham K, Vijayaraman TM (eds) Proceedings of the 19th international conference on data engineering, March 5–8, 2003, Bangalore, India, IEEE Computer Society, pp 315–326
Reimann C, Siewers U, Tarvainen T, Bityukova L, Eriksson J, Gilucis A, Gregorauskiene V, Lukashev VK, Matinian NN, Pasieczna A (2003) Agricultural soils in northern Europe: a geochemical atlas. In: Geologisches Jahrbuch. Schweizerbart'sche Verlagsbuchhandlung, Stuttgart
Riani M, Atkinson AC, Cerioli A (2009) Finding an unknown number of multivariate outliers. J R Stat Soc Ser B 71:447–466
Rousseeuw PJ, Leroy AM (2003) Robust regression and outlier detection. Wiley, New York
Rousseeuw PJ, Van Driessen K (1999) A fast algorithm for the minimum covariance determinant estimator. Technometrics 41(3):212–223
Rousseeuw PJ, Van Zomeren BC (1990) Unmasking multivariate outliers and leverage points. J Am Stat Assoc 85(411):633–651