Robust detection techniques for multivariate spatial data

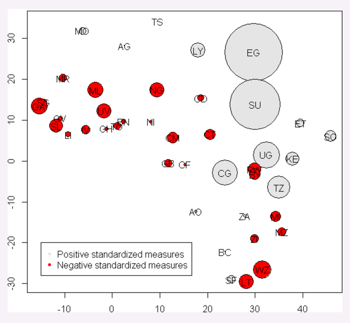
ERNST Marie   Department of Mathematics, University of Liege, Belgium   m.ernst@ulg.ac.be

Joint work with HAESBROECK Gentiane

Spatial data

Spatial data are characterized by n statistical units, with known geographical positions, on which p non spatial attributes are measured.

Example: A conflict measure in 42 african countries.

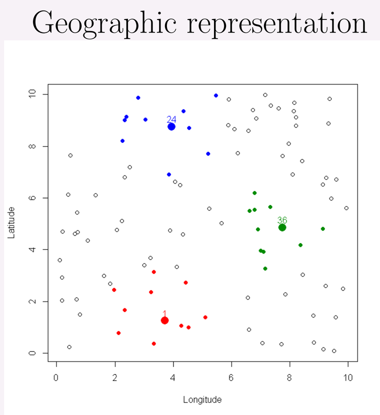


Spatial outlier

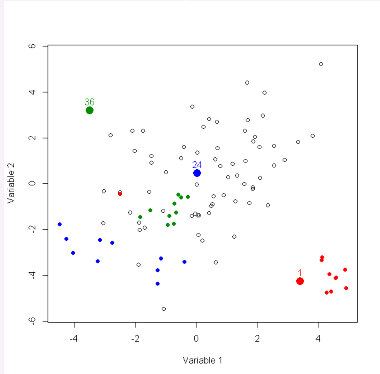
Haslett *et al.* [3] distinguishes two types of outliers in spatial data.

- A global outlier is an observation that might have non spatial attributes with significantly differing values wrt the majority of the data points.
- A local outlier is an observation that might have non spatial attributes with significantly differing values wrt its neighbors.

Geographic representation



Attribute representation



- The blue observation is a local but not global outlier.
- The green observation is a local and global outlier.
- The red observation is a global but not local outlier.

Covariance matrix estimator

- **Minimum Covariance Determinant (MCD) estimator**

$$S_H = \frac{1}{|H|}\sum_{i\in H}(x_i - \overline{x}_H)(x_i - \overline{x}_H)^T$$

for some specific subset H of $\{1,\dots,n\}$ that minimize the determinant. This estimator is robust but not invertible if $|H| < p$.

- **Regularized estimator**

$$(\hat{\mu}, \widehat{\Sigma}) = \underset{(\mu,\Sigma)}{\operatorname{argmax}} \left\{\log L(\mu,\Sigma) - \lambda J(\Sigma^{-1})\right\}$$

where J is a penalty function (*e.g.*, trace, L1 or L2 norm). The covariance matrix estimator is invertible.

- **Regularized MCD** [2]

$$(\hat{\mu}, \widehat{\Sigma}) = \underset{(\mu_H,\Sigma_H)}{\operatorname{argmax}} \left\{\log L(\mu_H,\Sigma_H) - \lambda J(\Sigma_H^{-1})\right\}$$

for the optimal subset H.

Detection technique of Filzmoser *et al.* [1]

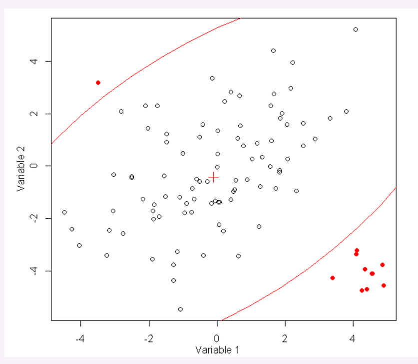
- **Global outlier detection :**

(a) Estimate robustly the general structure: MCD over the whole dataset gives $(\widehat{\mu},\widehat{\Sigma})$.

(b) Compute Mahalanobis distances between the center and each observation x_i $(i = 1,\dots,n)$:

$$MD_{(\widehat{\mu},\widehat{\Sigma})}(x_i) = (x_i - \widehat{\mu})^T\, \widehat{\Sigma}^{-1}\, (x_i - \widehat{\mu})$$

(c) If the distance $MD_{(\widehat{\mu},\widehat{\Sigma})}(x_i)$ is larger than a chisquare quantile then x_i is considered as a global outlier.



References :

[1] Filzmoser, P., Ruiz-Gazen, A. and Thomas-Agnan, C., Identification of Local Multivariate Outliers, *Statistical Papers*, (2013), 1–19.

[2] Fritsch, V., Varoquaux G., Thyreau, B., Poline, J.B. and Thirion, B., Detecting Outlying Subjects in High-Dimensional Neuroimaging Datasets with Regularized Minimum Covariance Determinant, *Medical Image Analysis*, **16**, (2012), 1359–1370.

[3] Haslett, J., Brandley, R., Craig, P., Unwin, A. and Wills, G., Dynamic Graphics for Exploring Spatial Data With Applications to Locating Global and Local Anomalies, *The American Statistician*, **45**, (1991), 234–242.

[4] Paindaveine, D. and Van Bever, G., From Depth to Local Depth : a Focus on Centrality, *Journal of the American Statistical Association*, **105**, (2013), 1105–1119.

[5] Zuo, Y. and Serfling, R., General Notions of Statistical Depth Function, *The Annals of Statistics*, **28**, (2000), 461–482.

Detection technique of Filzmoser *et al.* [1]

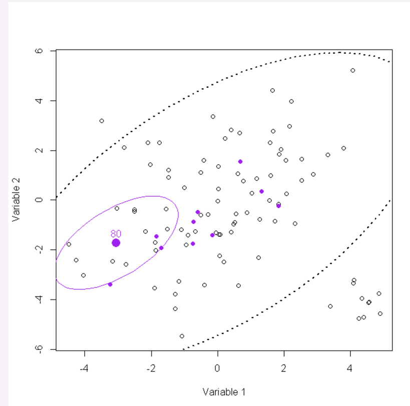
- **Local outlier detection :**

For each observation x_i $(i = 1,\dots,n)$:

(a) Compute the pairwise Mahalanobis distances between x_i and its k neighbors x_j using the global structure:
$MD_{\widehat{\Sigma}}(x_i,x_j) = (x_i - x_j)^T\, \widehat{\Sigma}^{-1}\, (x_i - x_j)$.

(b) Determine the ellipsoid containing a proportion β of its k neighbors.

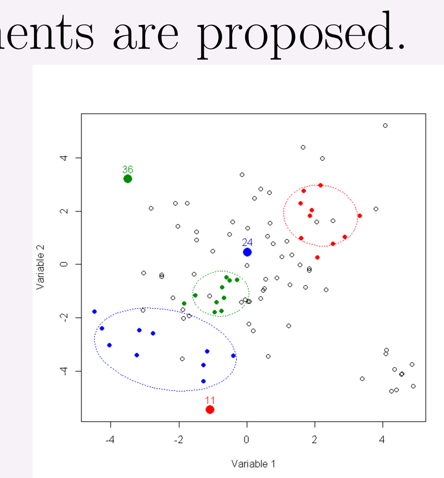
(c) If the tolerance level of this ellipsoid is too large according to the chisquare distribution then the observation is considered as a local outlier.



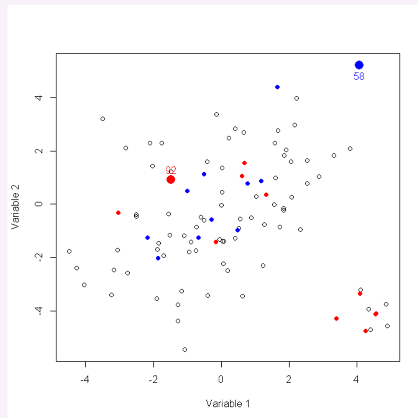
Proposition 1 : parametric technique

This proposition is an adaptation of the technique presented by Filzmoser *et al.* [1] for the local outlier detection. Two improvements are proposed.

1. Use a local structure estimated separately on each neighborhood instead of the general one.
As the size k of the neighborhood can be smaller than the dimension p, the local structure has to be estimated by a robust and regularized estimator.



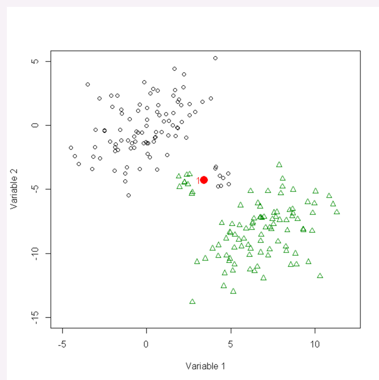
2. Instead of testing the local outlyingness of each observation, we suggest to focus only on the observations corresponding to a positively spatially autocorrelated neighborhood.
The multivariate autocorrelation of a neighborhood is estimated by means of the determinant of the regularized MCD covariance estimator computed on the neighborhood and only the neighborhoods yeilding the smallest values are selected.



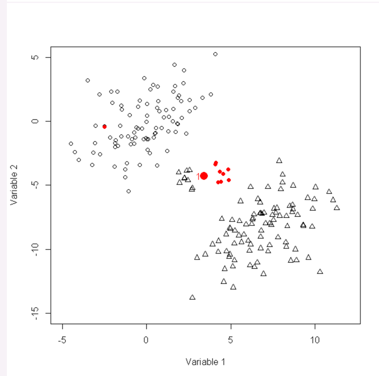
Proposition 2 : non parametric technique

This non parametric detection technique for local outliers is based on depth functions [5].
As in the first proposition, local outlyingness is tested only on positively spatially autocorrelated neighborhoods. By definition the neighbors of a local outlier are "far" from it according to other observations.

To compare an observation x_i and its neighbors, let's make x_i the deepest point (the center) by using the symmetrized dataset [4]. Then calculate the depth values of its neighbors in this new dataset.



If the $(\beta k)^{th}$ depth is too small or equivalently, if more than a proportion β of its neighbors are too far according to other observations then x_i is considered as a local outlier.



On going research

Some partial findings are:

- Restricting the detection to the positively spatially autocorrelated neighborhoods is necessary to avoid increasing the "false-positive" detection rate;
- The chisquare distribution is not a good approximation for the distribution of the "regularized" robust distances;
- The tuning of the parameters (k, β) still needs to be improved.