Comparison of robust detection techniques for local outliers in multivariate spatial data

Marie Ernst and Gentiane Haesbroeck

University of Liege

COMPSTAT 2014 – Geneva

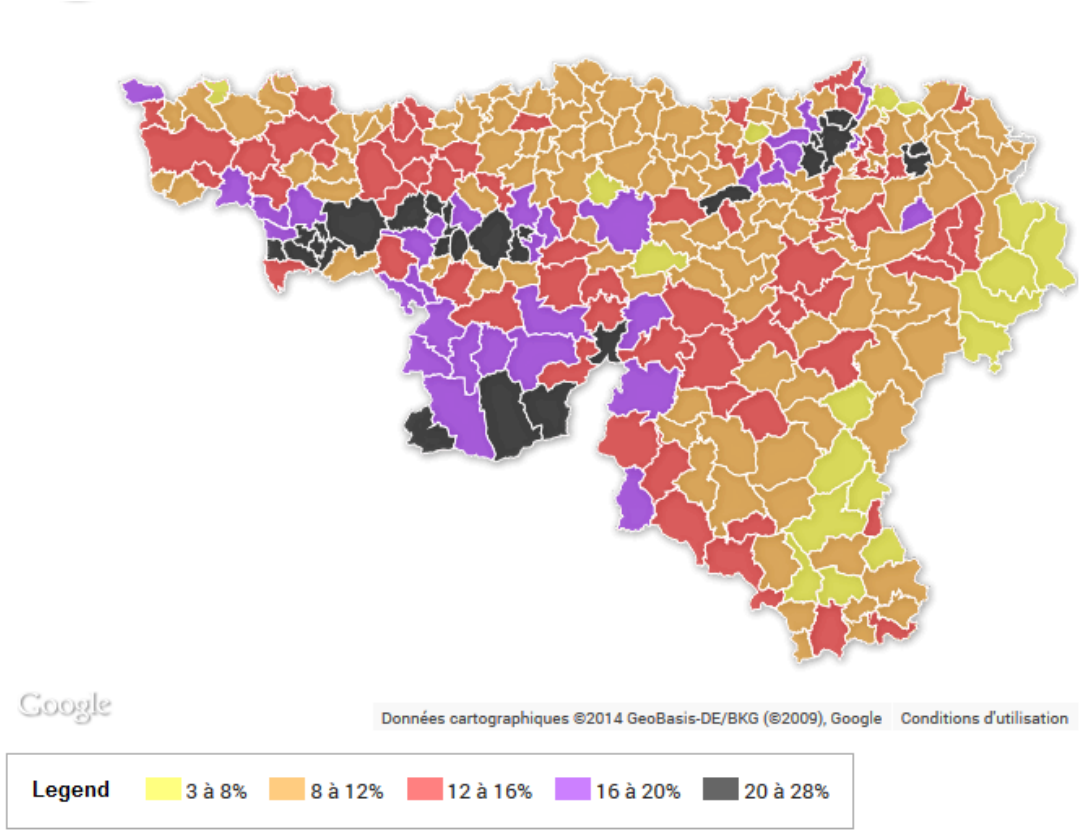
Spatial Data

Spatial data :

- geographical positions
- non spatial attributes

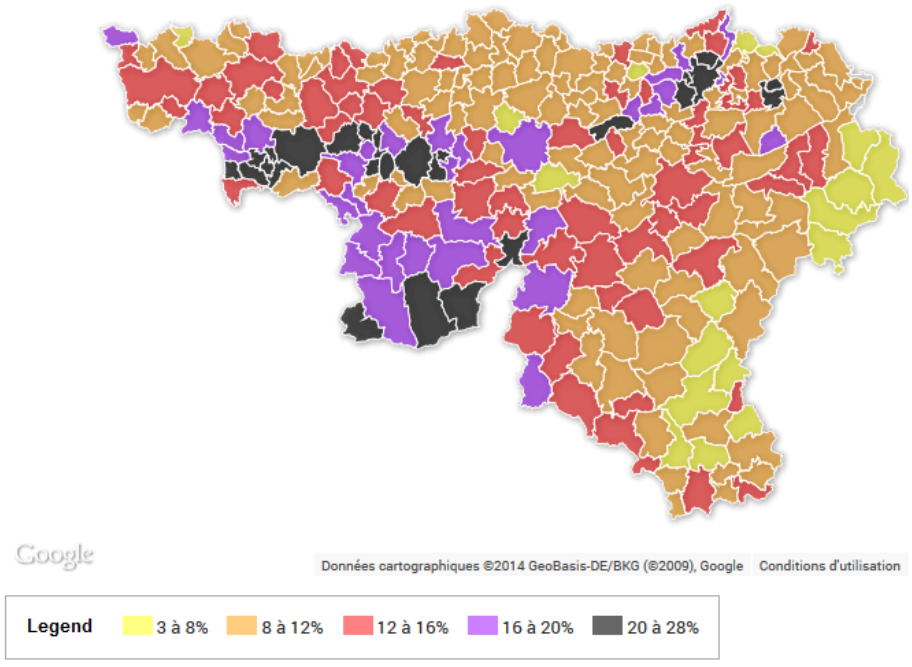
Example 1

Unemployment rate in the Walloon region in Belgium

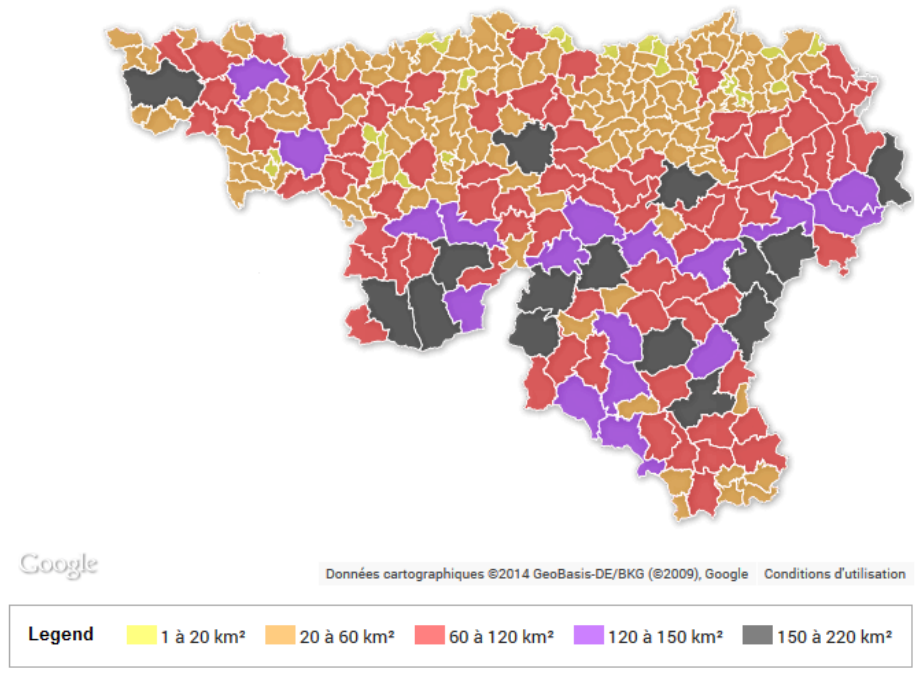


Example 2

Unemployment rate

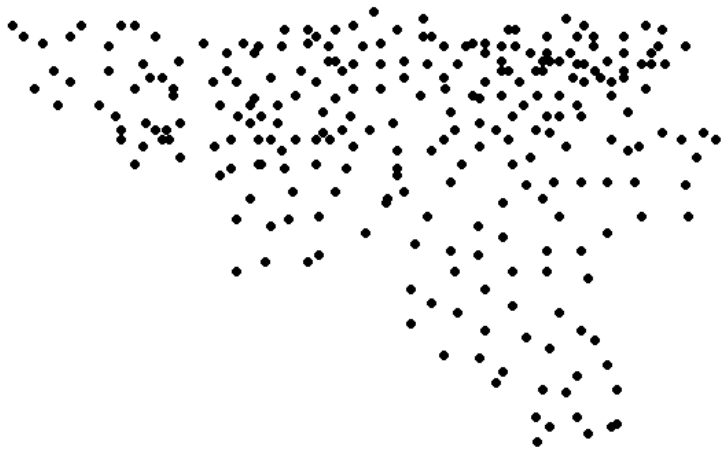


Surface area

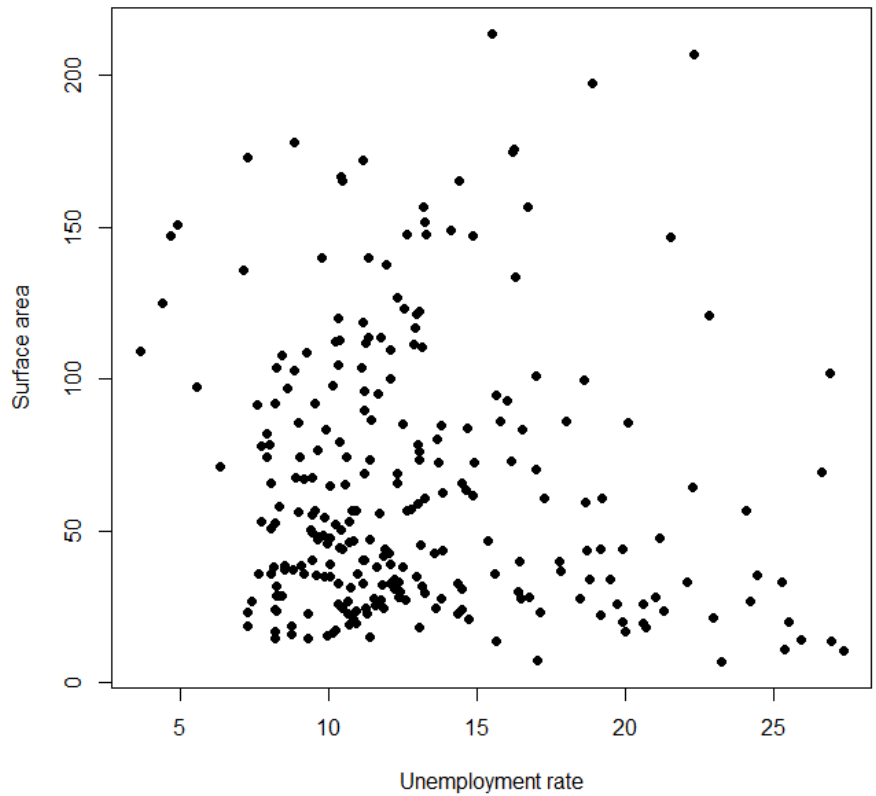


Example 2

Spatial coordinates
(Latitude/Longitude)



Unemployment rate
vs Surface area

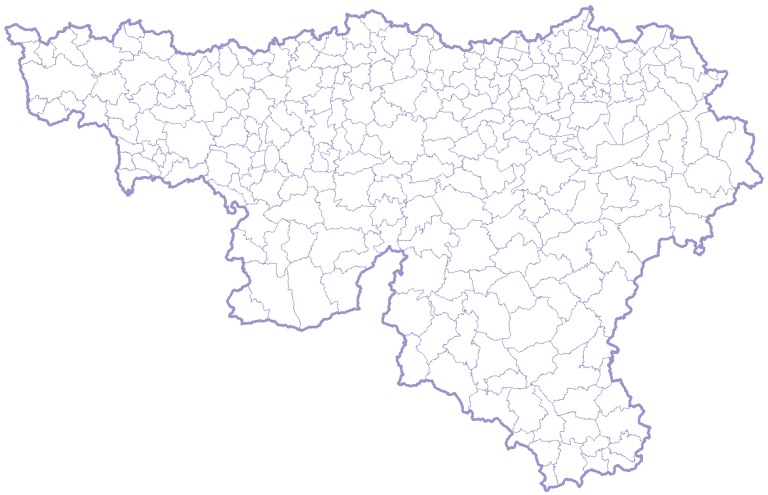


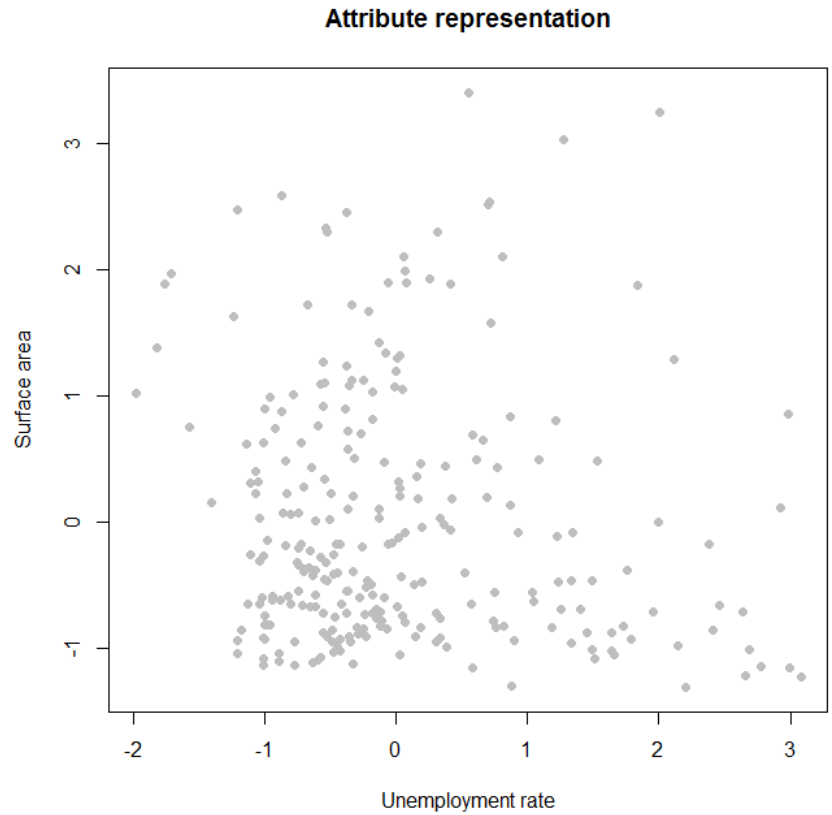
Outlier definitions

Two types of outliers (Haslett *et al.* (1991)) :

- local outlier : extreme behavior wrt neighbors
- global outlier : extreme behavior wrt all observations

Example 2



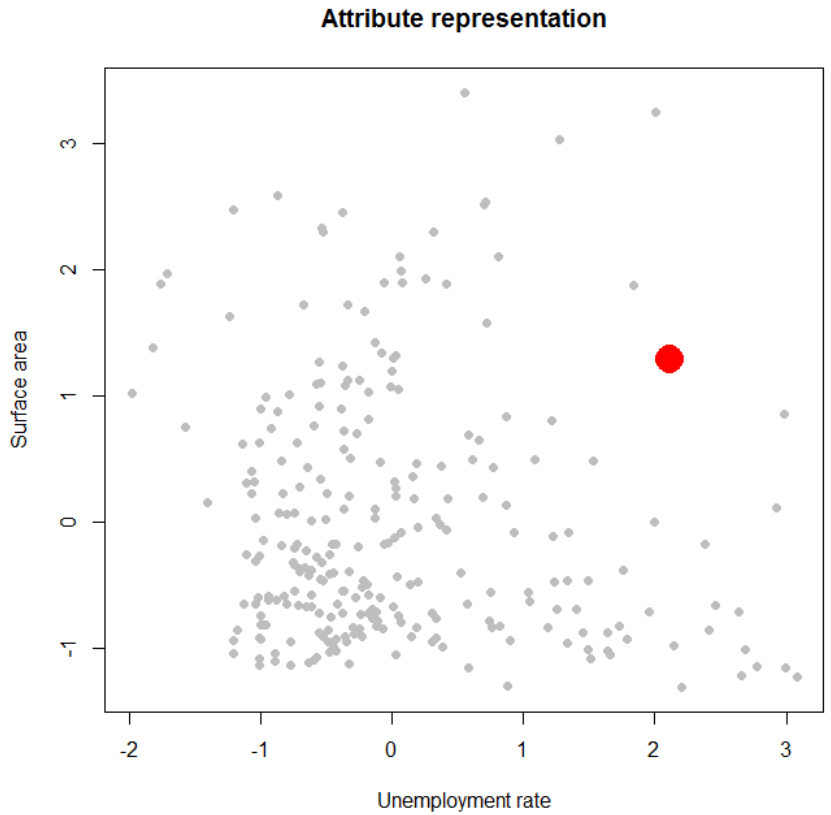


Outlier definitions

Two types of outliers (Haslett *et al.* (1991)) :

- local outlier : extreme behavior wrt neighbors
- global outlier : extreme behavior wrt all observations

Example 2

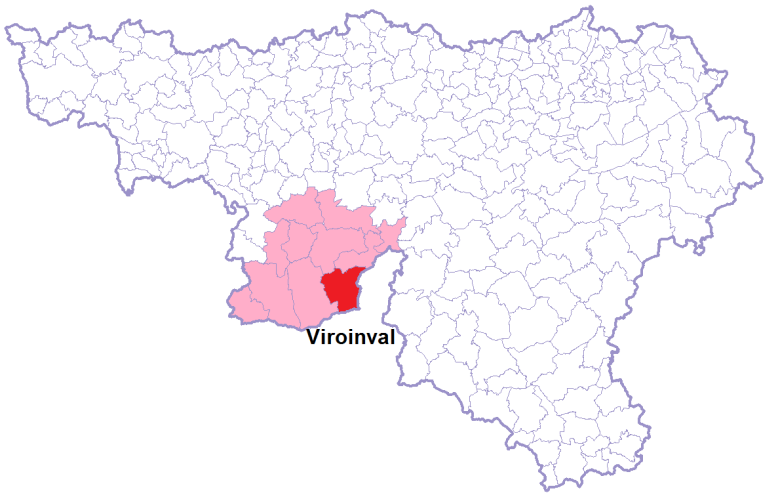


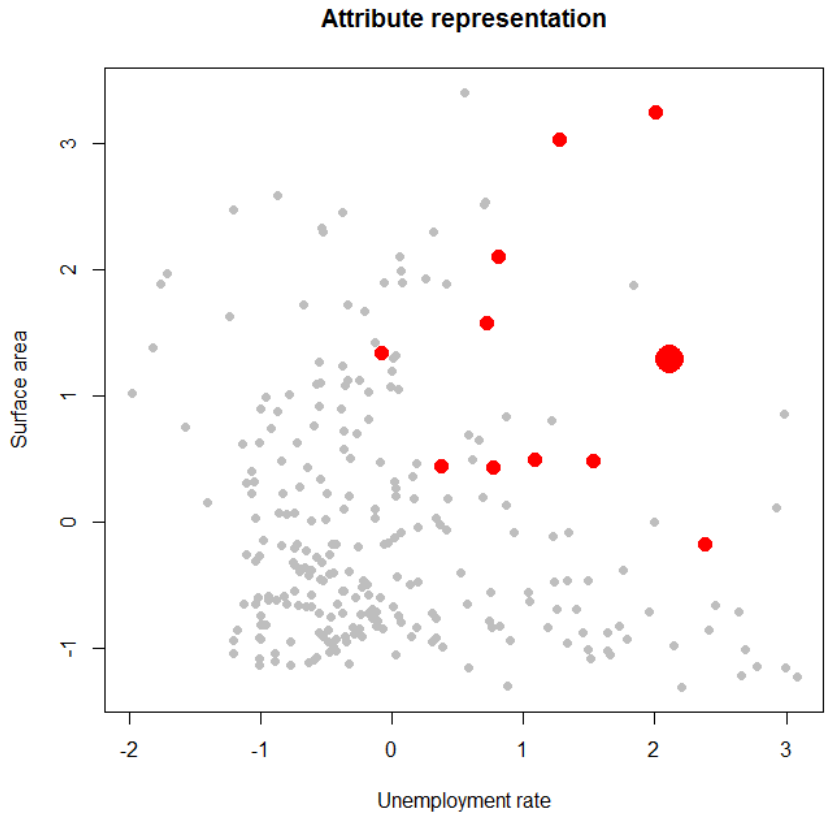
Outlier definitions

Two types of outliers (Haslett *et al.* (1991)) :

- local outlier : extreme behavior wrt neighbors
- global outlier : extreme behavior wrt all observations

Example 2



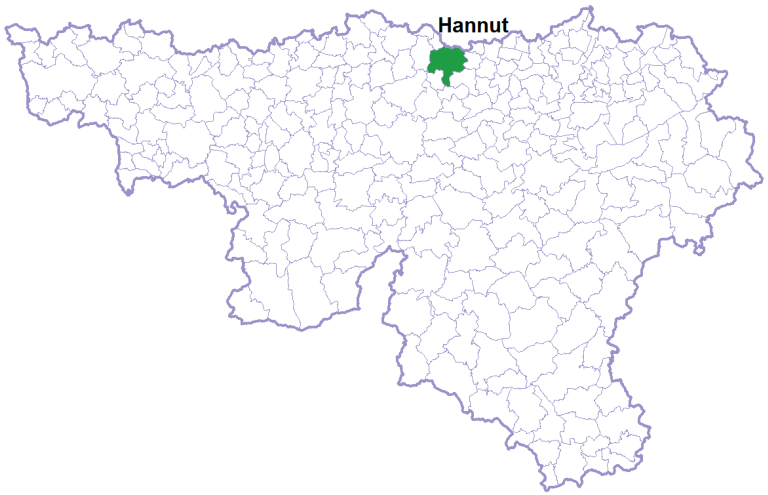


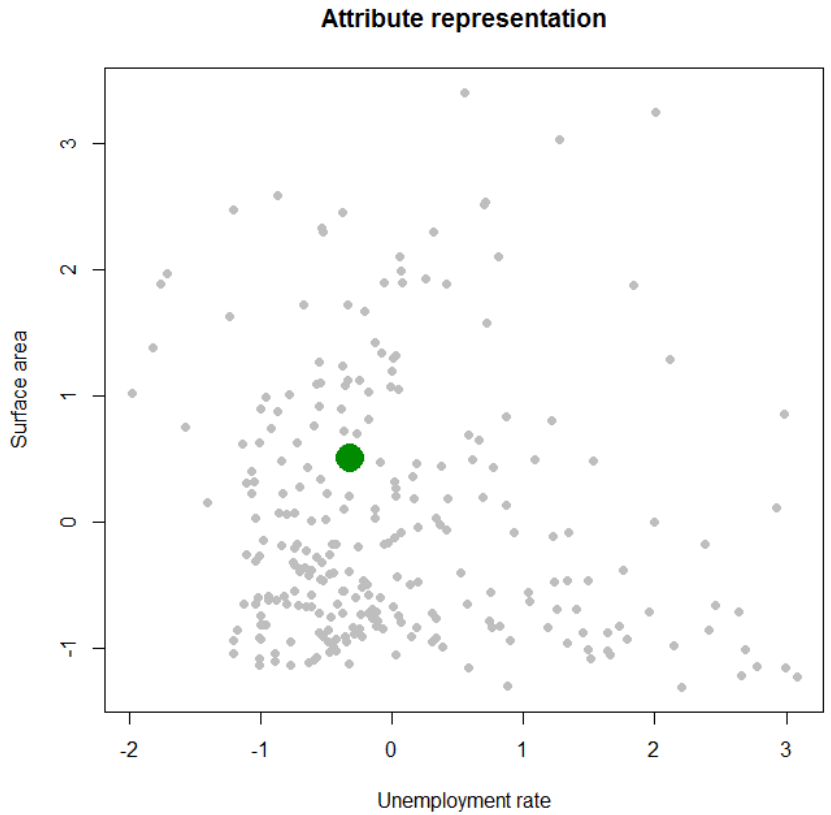
Outlier definitions

Two types of outliers (Haslett *et al.* (1991)) :

- local outlier : extreme behavior wrt neighbors
- global outlier : extreme behavior wrt all observations

Example 2



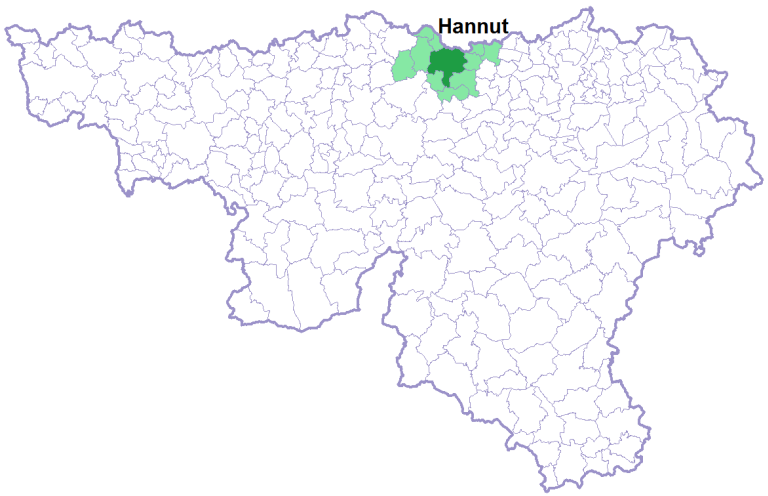


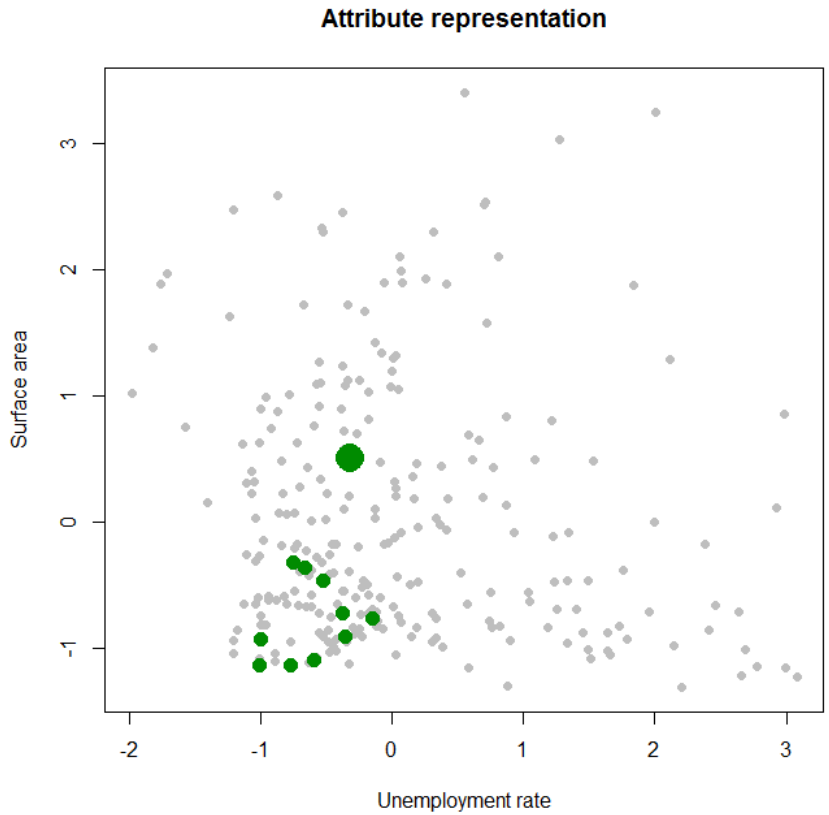
Outlier definitions

Two types of outliers (Haslett *et al.* (1991)) :

- local outlier : extreme behavior wrt neighbors
- global outlier : extreme behavior wrt all observations

Example 2



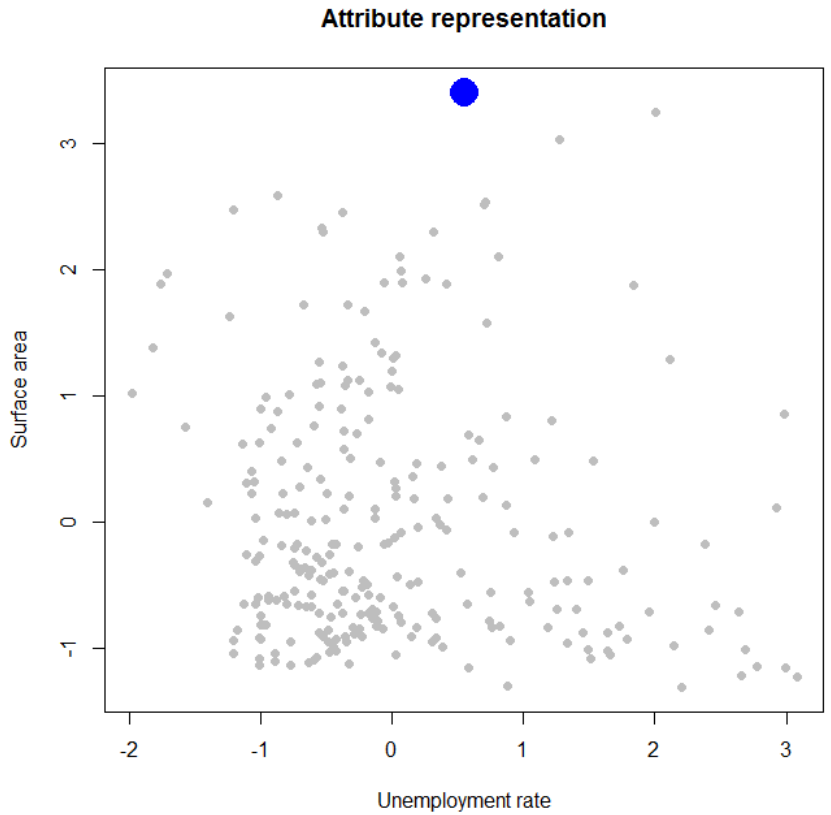


Outlier definitions

Two types of outliers (Haslett *et al.* (1991)) :

- local outlier : extreme behavior wrt neighbors
- global outlier : extreme behavior wrt all observations

Example 2

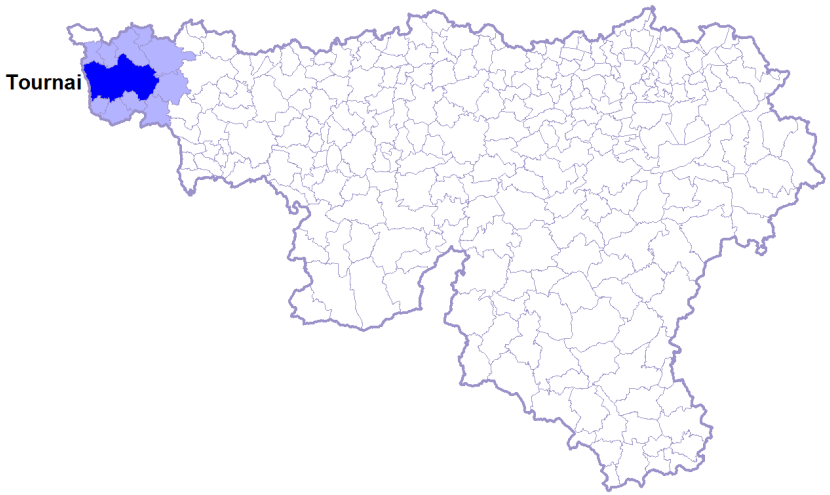


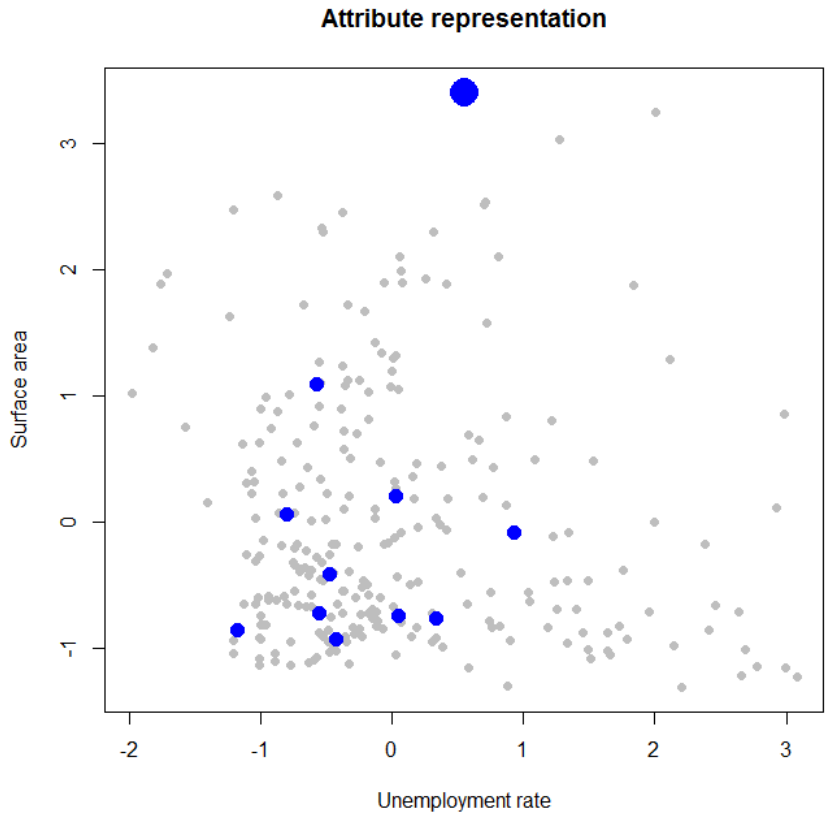
Outlier definitions

Two types of outliers (Haslett *et al.* (1991)) :

- local outlier : extreme behavior wrt neighbors
- global outlier : extreme behavior wrt all observations

Example 2





Objectives in dimension p

Global outliers detection

- Geographical components not used
- Usual outlier detection techniques can be used $\Rightarrow$ not considered here

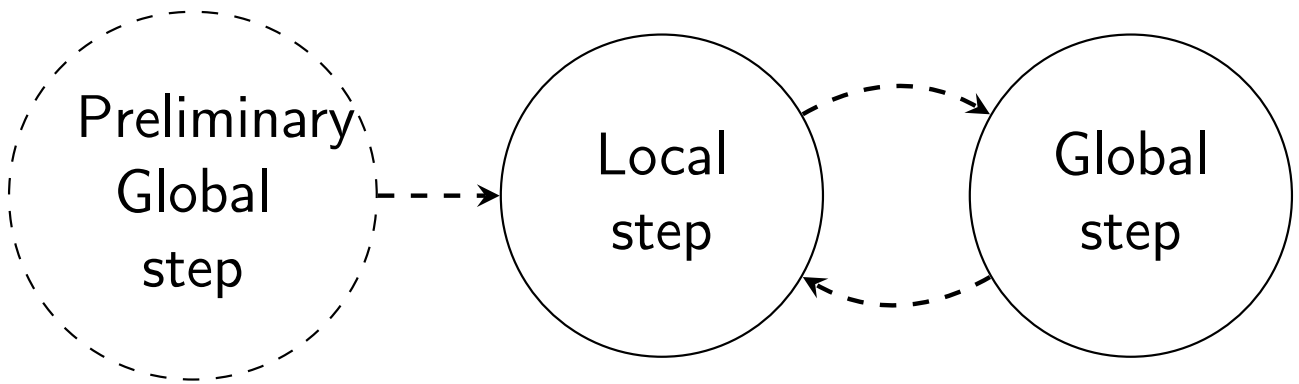
Local outliers detection

- Review of some existing techniques
- Suggestion of an adaptation
- Comparison on examples and simulations

Considered Techniques

- Chen *et al.* (2008)
- Harris *et al.* (2014)
- Filzmoser *et al.* (2014)
- A new proposal: regularized version of Filzmoser *et al.* (2014)

Review of Shubert *et al.* (2014)



Chen *et al.* (2008)

Approach

Using Componentwise median and robust Mahalanobis distances

1. Preliminary global step: standardization
2. Local step: $h_i = z_i - g(z_i)$, for an observation $z_i \in \mathbb{R}^p$ and the componentwise median $g(z_i)$ over its neighborhood
3. Global step: work on $\{h_1, \ldots, h_n\}$
   - Robust estimation of the general structure: $(\hat{\mu}, \hat{\Sigma})$
   - Mahalanobis distances: $MD_{(\hat{\mu},\hat{\Sigma})}(h_i) = (h_i - \hat{\mu})' \hat{\Sigma}^{-1} (h_i - \hat{\mu})$
   - If the distance is larger than a predefined threshold $\Rightarrow$ **local outlier**

Harris *et al.* (2014)

Approach

Using Geographically Weighted PCA with robust estimator

1. Preliminary global step:
   If the dimension is too large $\Rightarrow$ Robust PCA to retain q components
2. Local step:
   - Local PCA on each neighborhood
   - Comparison of local score distances with a theoretical quantile
   - Comparison of orthogonal distances with empirical quantiles
   - Comparison of PCA scores with empirical quantiles

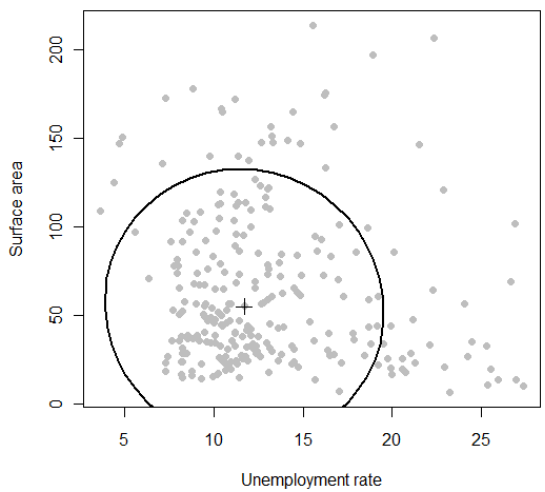
Filzmoser *et al.* (2014)

Approach

Robust "Mahalanobis-type" detection

1. **Preliminary global step:**
   Robust estimation of the general structure: $(\hat{\mu}, \hat{\Sigma})$
2. Local step:
   - Centring the general structure on the observation
   - Determination of the ellipsoid containing the next neighbor
   - If its tolerance level is larger than a theoretical quantile $\Rightarrow$ **local outlier**

Example 2



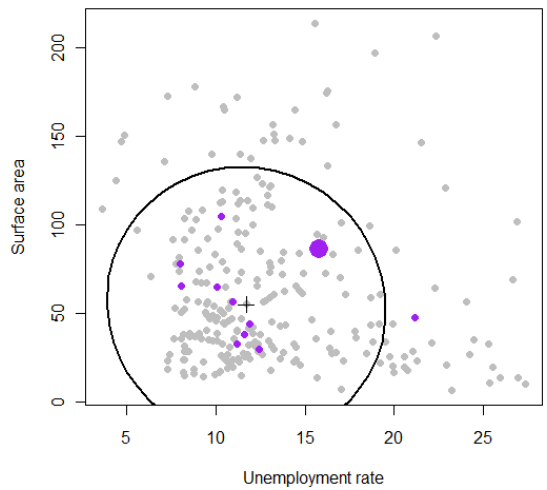
Filzmoser *et al.* (2014)

Approach

Robust "Mahalanobis-type" detection

1. **Preliminary global step:**
   Robust estimation of the general structure: $(\hat{\mu}, \hat{\Sigma})$
2. **Local step:**
   - Centring the general structure on the observation
   - Determination of the ellipsoid containing the next neighbor
   - If its tolerance level is larger than a theoretical quantile $\Rightarrow$ **local outlier**

Example 2



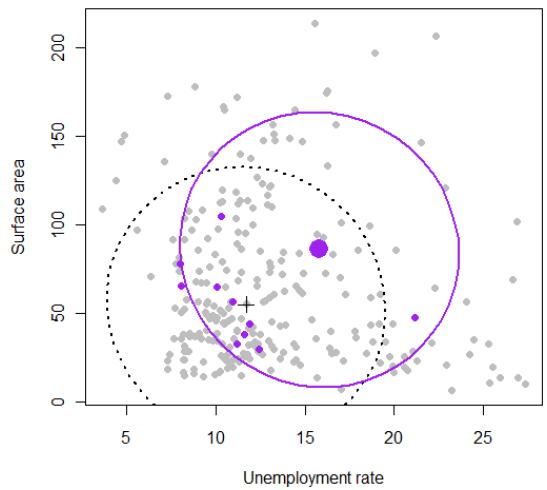
Filzmoser *et al.* (2014)

Approach

Robust "Mahalanobis-type" detection

1. **Preliminary global step:**
   Robust estimation of the general structure: $(\hat{\mu}, \hat{\Sigma})$
2. **Local step:**
   - Centring the general structure on the observation
   - Determination of the ellipsoid containing the next neighbor
   - If its tolerance level is larger than a theoretical quantile $\Rightarrow$ **local outlier**

Example 2



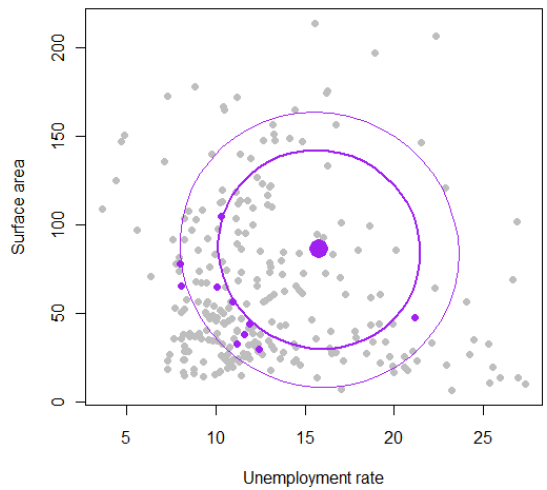
Filzmoser *et al.* (2014)

Approach

Robust "Mahalanobis-type" detection

1. Preliminary global step:
   Robust estimation of the general structure: $(\hat{\mu}, \hat{\Sigma})$
2. Local step:
   - Centring the general structure on the observation
   - Determination of the ellipsoid containing the next neighbor
   - If its tolerance level is larger than a theoretical quantile $\Rightarrow$ **local outlier**

Example 2



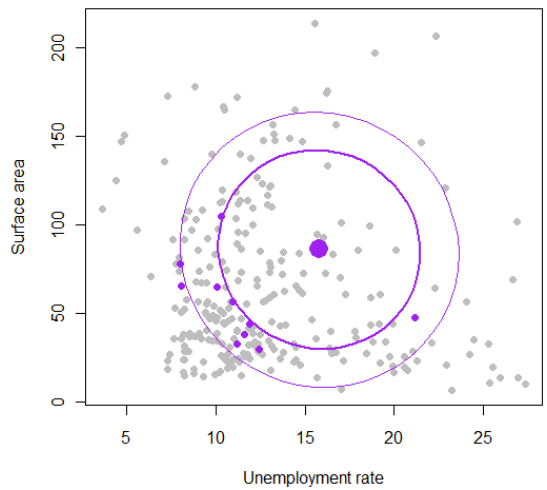
Filzmoser *et al.* (2014)

Approach

Robust "Mahalanobis-type" detection

1. Preliminary global step:
   Robust estimation of the general structure: $(\hat{\mu}, \hat{\Sigma})$
2. Local step:
   - Centring the general structure on the observation
   - Determination of the ellipsoid containing the next neighbor
   - If its tolerance level is larger than a theoretical quantile $\Rightarrow$ **local outlier**

Example 2



Regularized Filzmoser

Approach: adaptation of Filzmoser *et al.* (2014)

Work with local structure and only on most homogeneous neighborhoods

1. Local step:
   - Estimation of the local structure: $(\hat{\mu}_i, \hat{\Sigma}_i)$
   - Homogeneity measure: $\det(\hat{\Sigma}_i)$
2. Global step: Selection of 10%, 20%, 30% or 40% of smallest values
3. Local step: work only on *selected* neighborhoods
   - Centring the *local* structure on the observation
   - Determination of the ellipsoid containing the next neighbor
   - If its tolerance level is larger than an empirical quantile $\Rightarrow$ **local outlier**

Local structure

Robust and regularized estimator

Robust and regularized estimator

Robust estimator : MCD

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

for some specific subset H of $\{1, \ldots, n\}$. $\Rightarrow$ **Not invertible if** $|H| < p$

Regularized estimator

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

where J is a penalty function.

Regularized MCD[1]

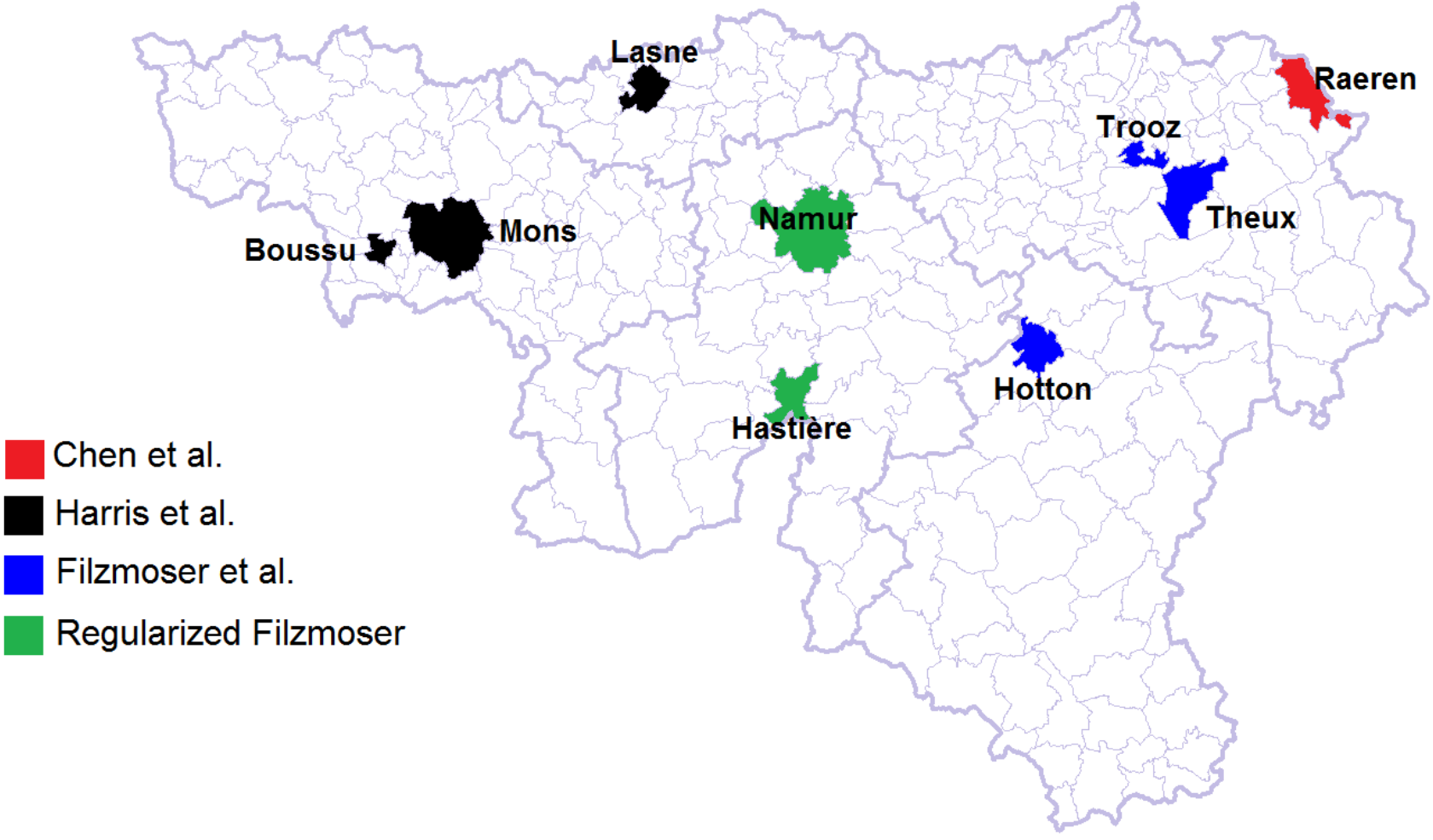
$$(\hat{\mu}, \widehat{\Sigma}) = \underset{(\mu_H, \Sigma_H)}{\text{argmax}} \left\{ \log L(\mu_H, \Sigma_H) - \lambda J(\Sigma_H^{-1}) \right\}$$

for the optimal subset H.

[1] Fritsch *et al* (2011).

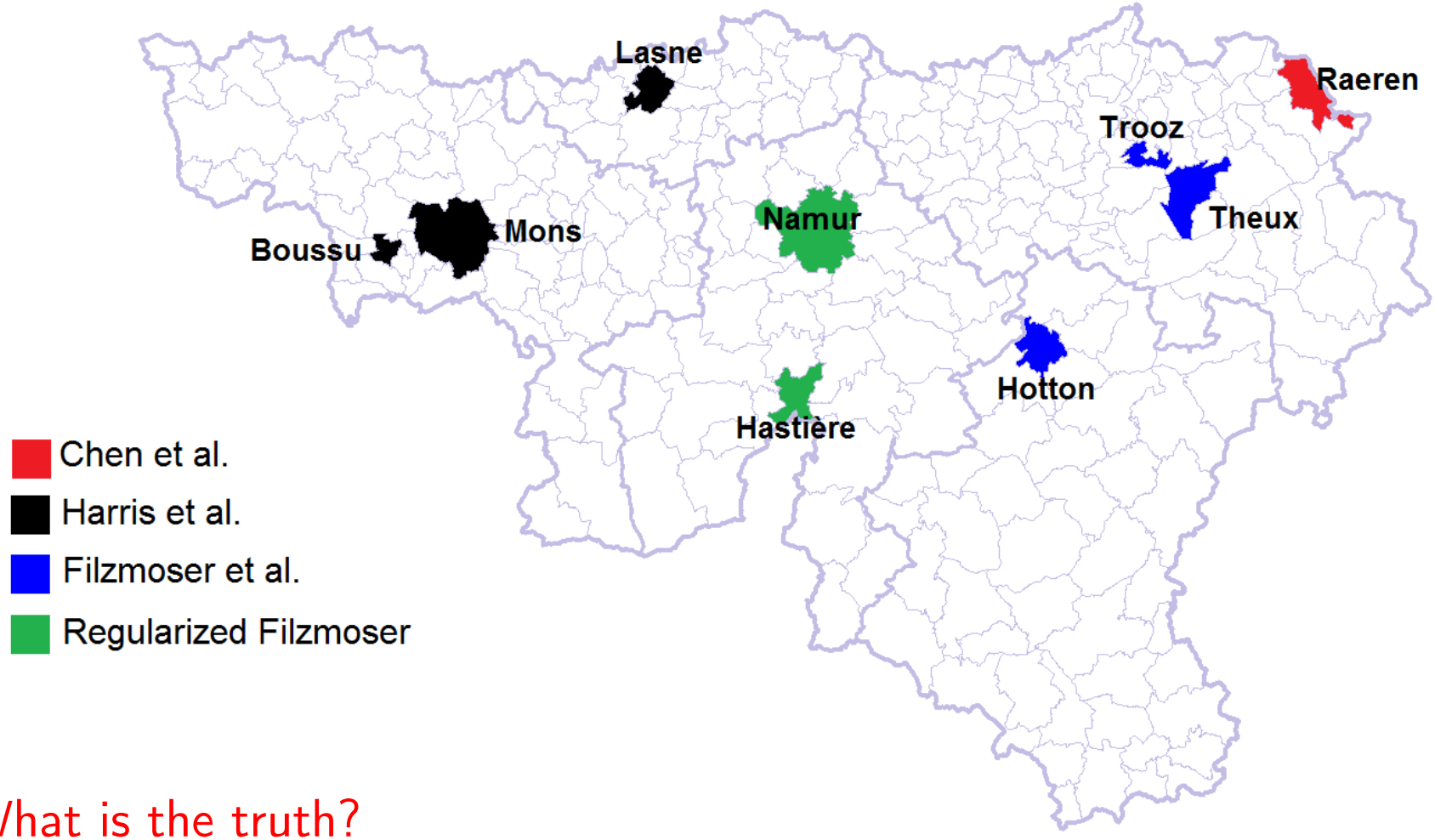
Illustrations

Wallonia: 14 variables for the 262 municipalities



Illustrations

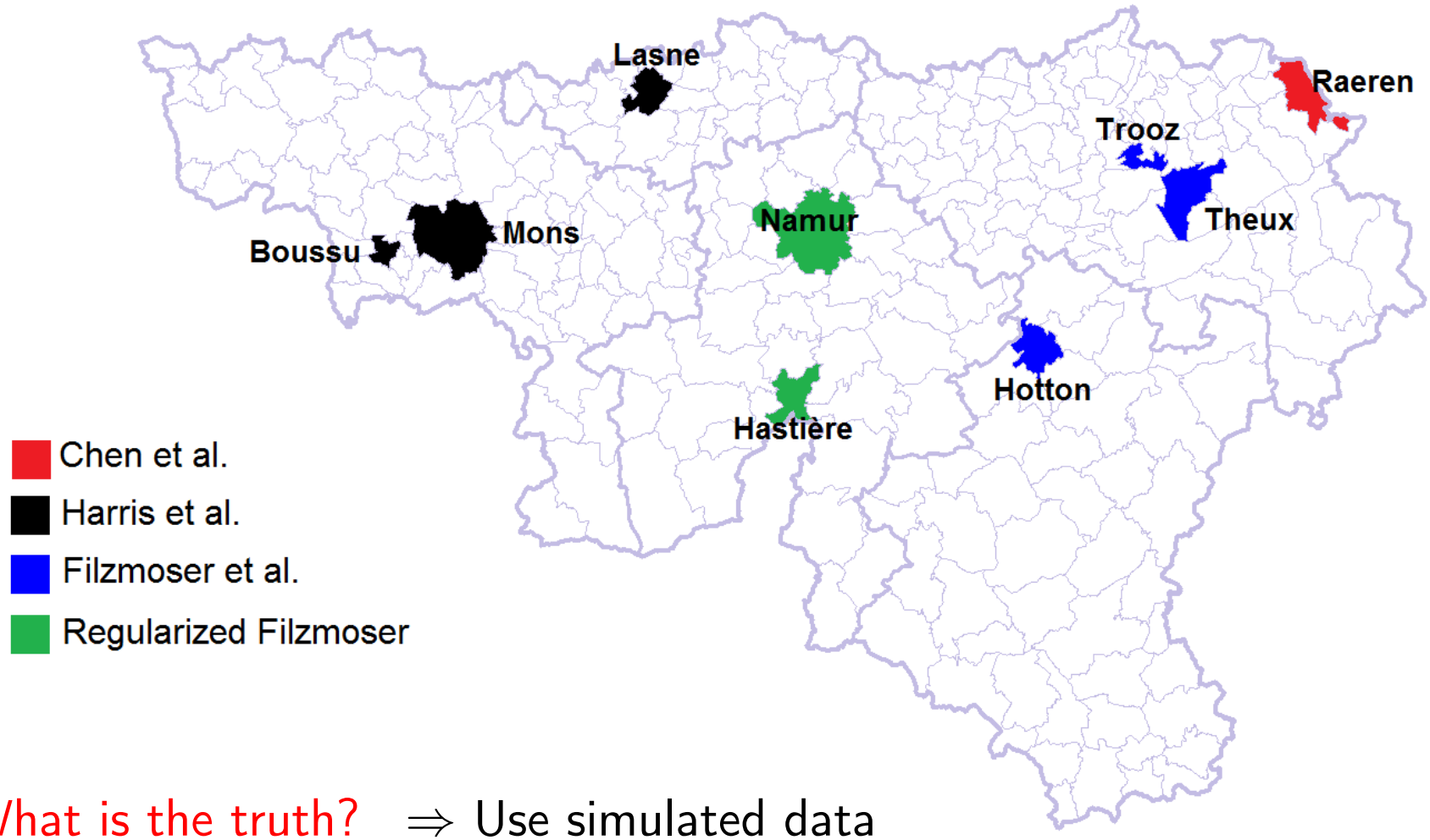
Wallonia: 14 variables for the 262 municipalities



What is the truth?

Illustrations

Wallonia: 14 variables for the 262 municipalities

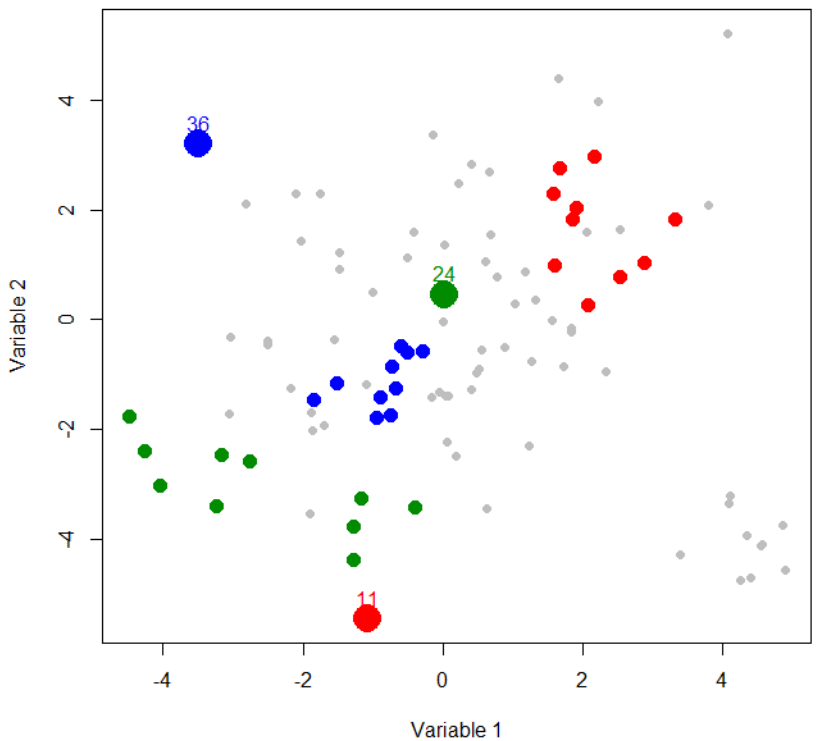


What is the truth? $\Rightarrow$ Use simulated data

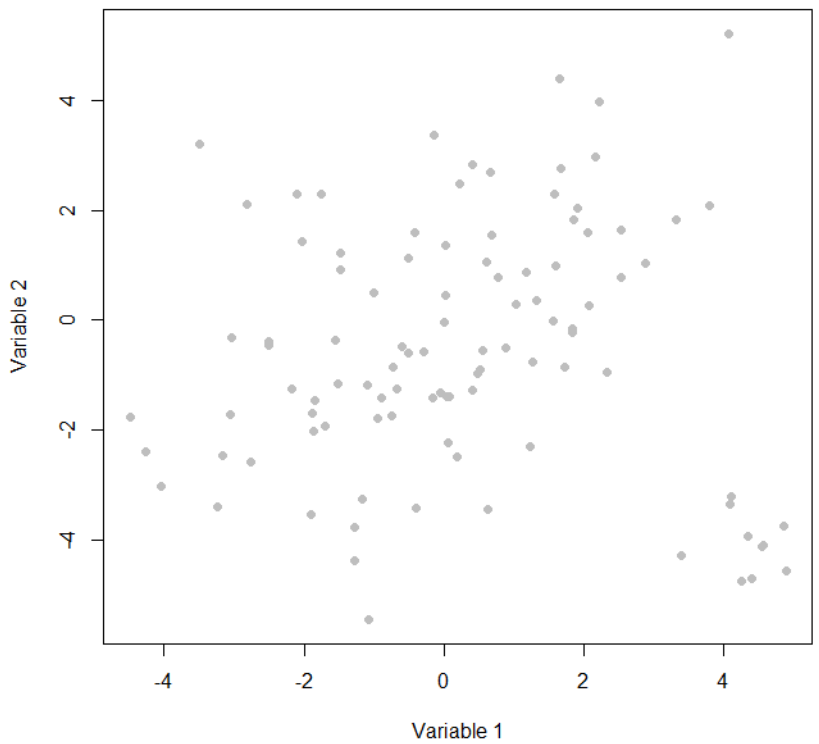
Illustration

Example 3 (Simulated data)[2]

Outliers for Filzmoser *et al.*



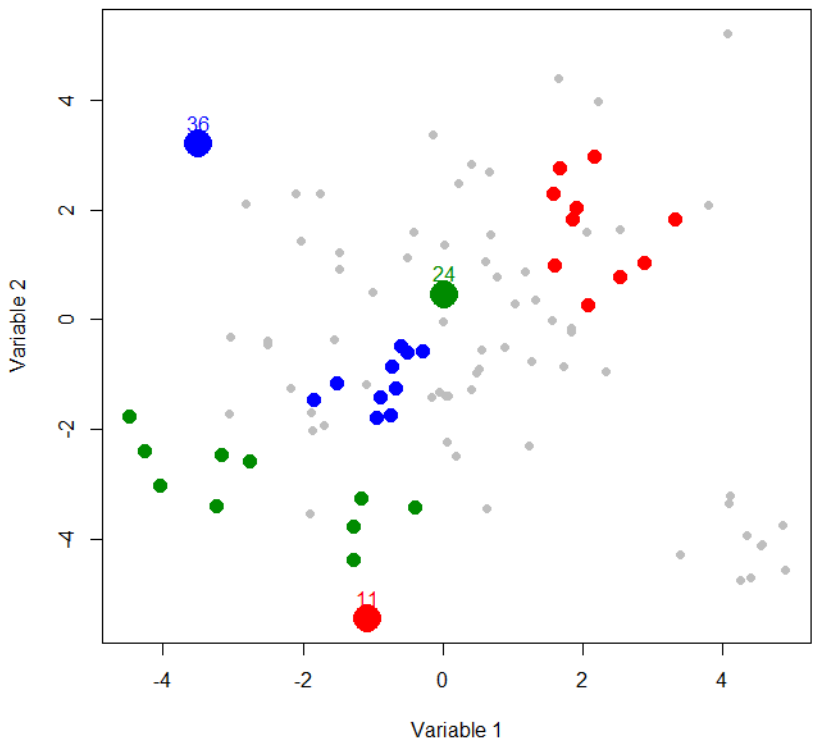
Outliers for the regularization



[2] Package mvoutlier in R, P. Filzmoser *et al.* (2014)

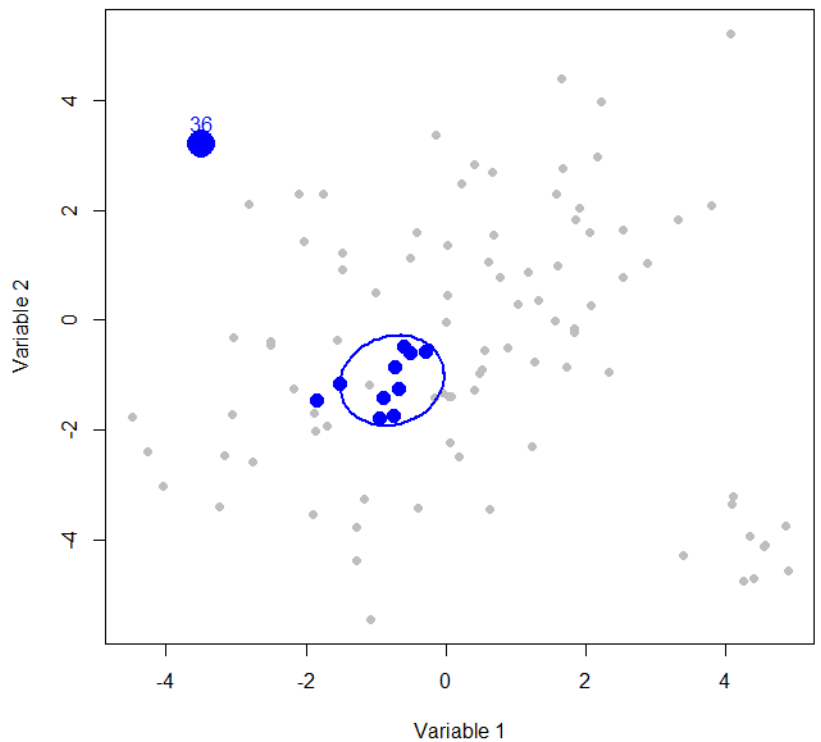
Example 3 (Simulated data)[2]

Outliers for Filzmoser *et al.*



Outliers for the regularization

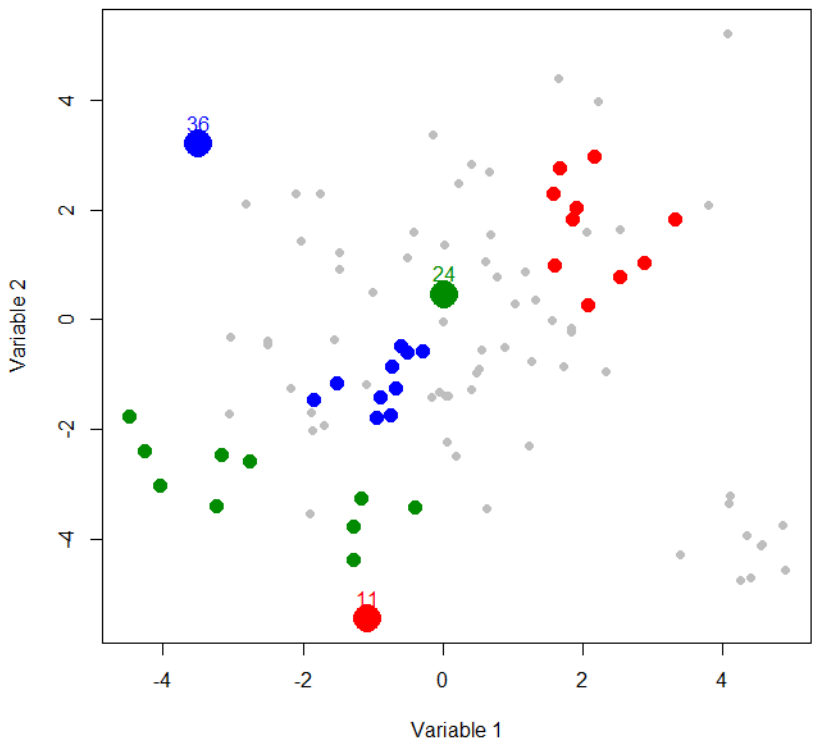
Test on 10% of neighborhoods



[2] Package mvoutlier in R, P. Filzmoser *et al.* (2014)

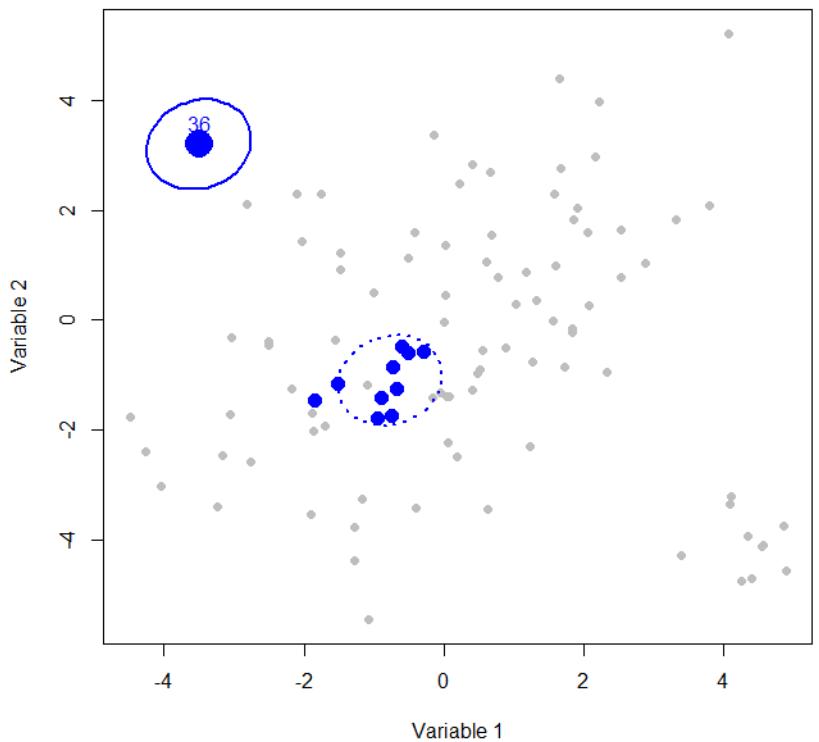
Example 3 (Simulated data)[2]

Outliers for Filzmoser *et al.*



Outliers for the regularization

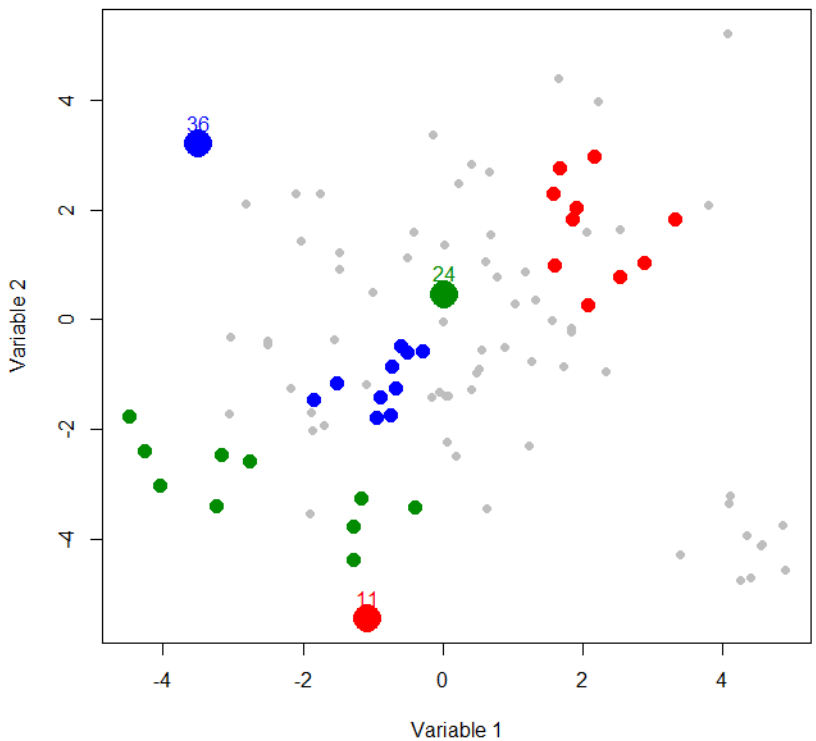
Test on 10% of neighborhoods



[2] Package mvoutlier in R, P. Filzmoser *et al.* (2014)

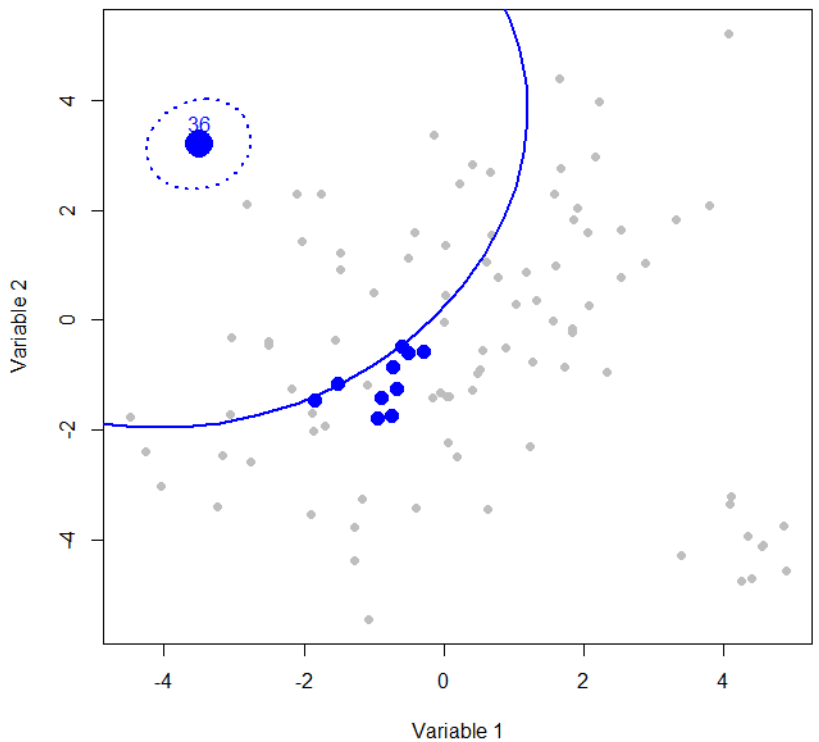
Example 3 (Simulated data)[2]

Outliers for Filzmoser *et al.*



Outliers for the regularization

Test on 10% of neighborhoods

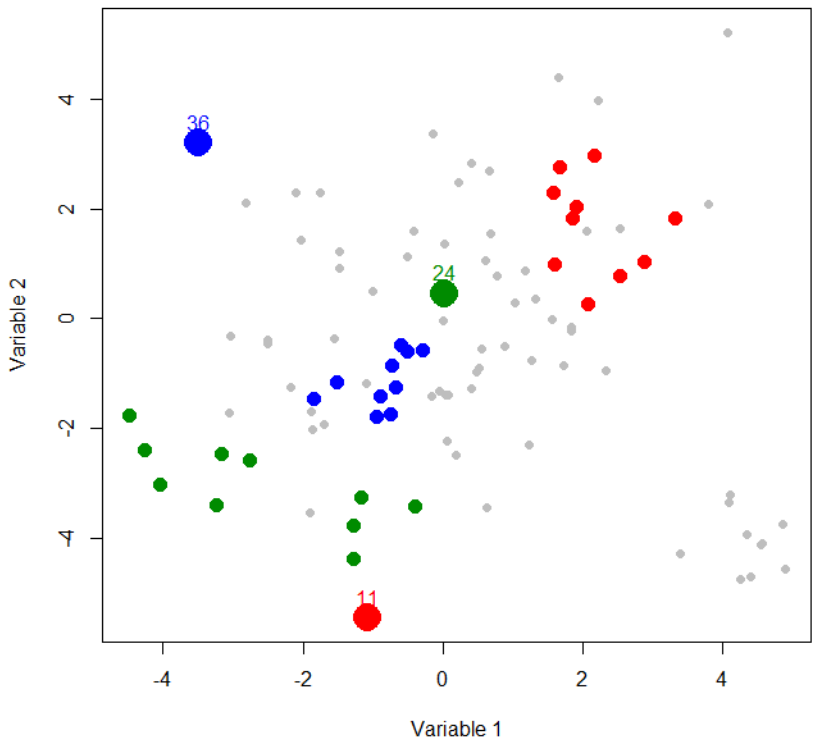


[2] Package mvoutlier in R, P. Filzmoser *et al.* (2014)

Illustration

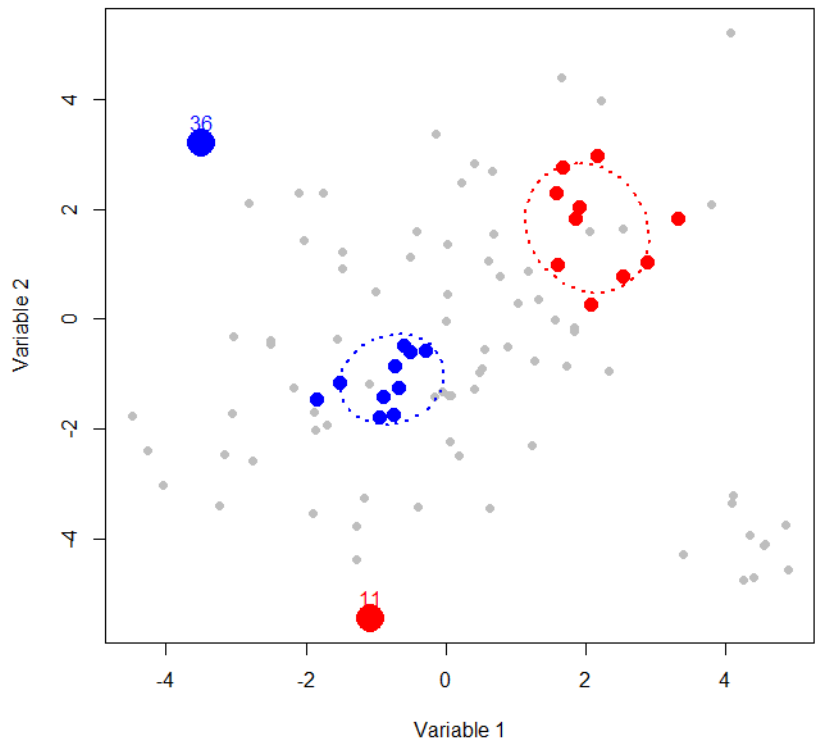
Example 3 (Simulated data)[2]

Outliers for Filzmoser *et al.*



Outliers for the regularization

Test on 20% of neighborhoods

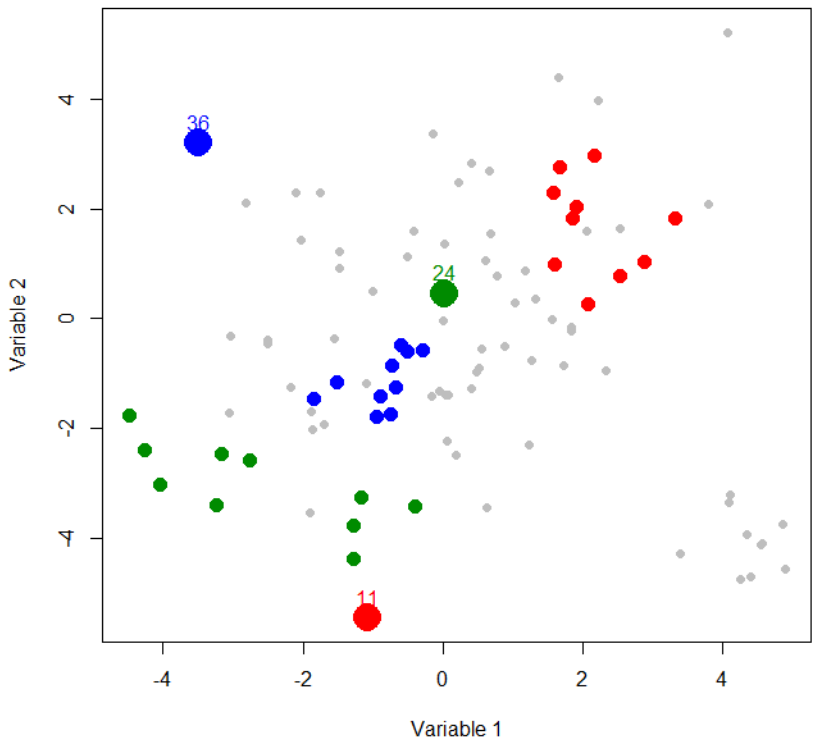


[2] Package mvoutlier in R, P. Filzmoser *et al.* (2014)

Illustration

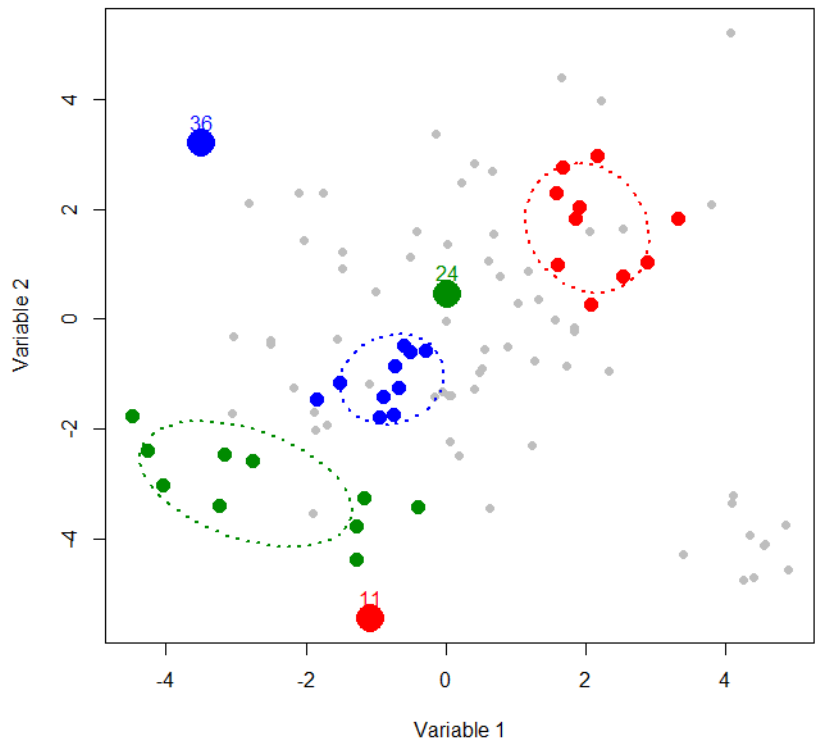
Example 3 (Simulated data)[2]

Outliers for Filzmoser *et al.*



Outliers for the regularization

Test on 30% of neighborhoods

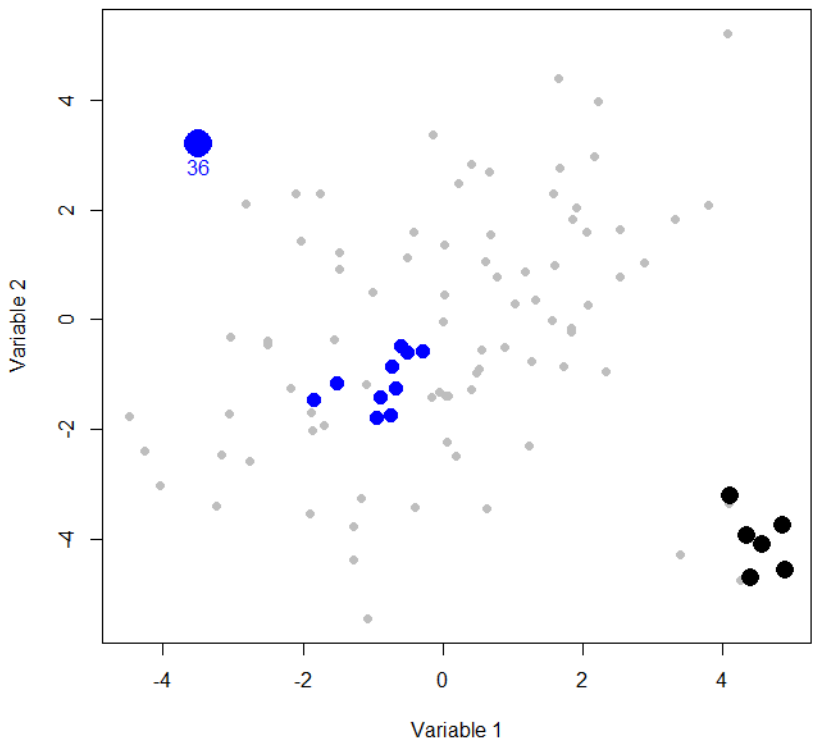


[2] Package mvoutlier in R, P. Filzmoser *et al.* (2014)

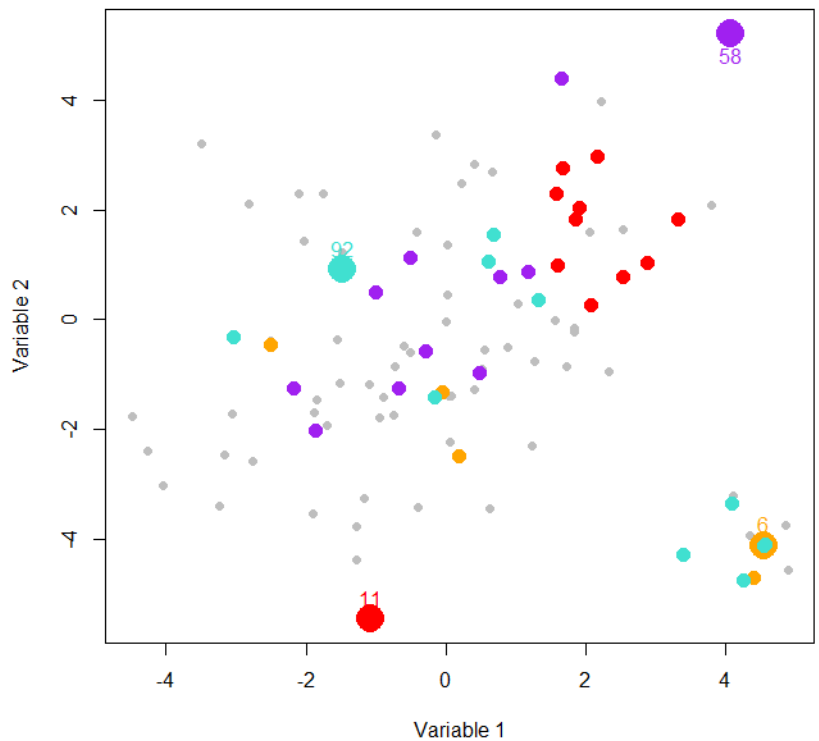
Illustration

Example 3 (Simulated data)[2]

Outliers for Harris *et al.*



Outliers for Chen *et al.*



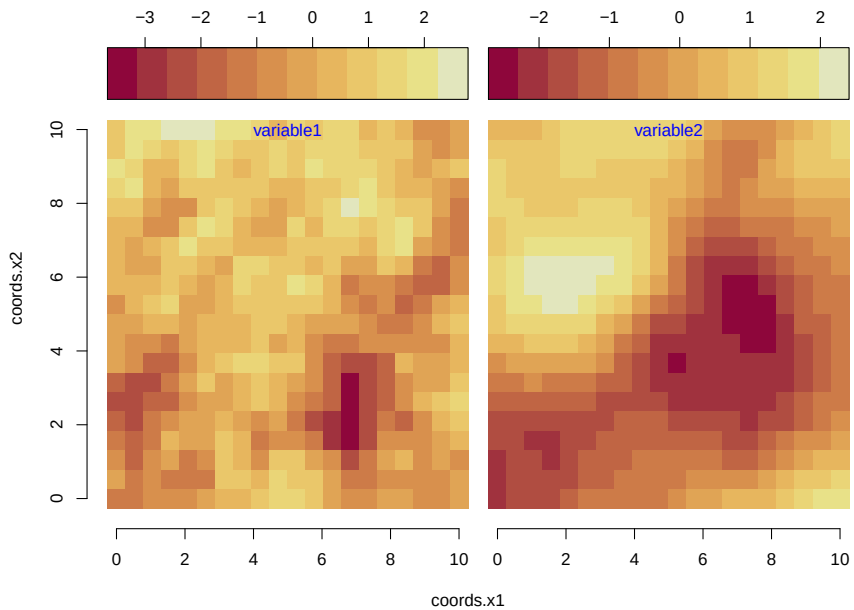
[2] Package mvoutlier in R, P. Filzmoser *et al.* (2014)

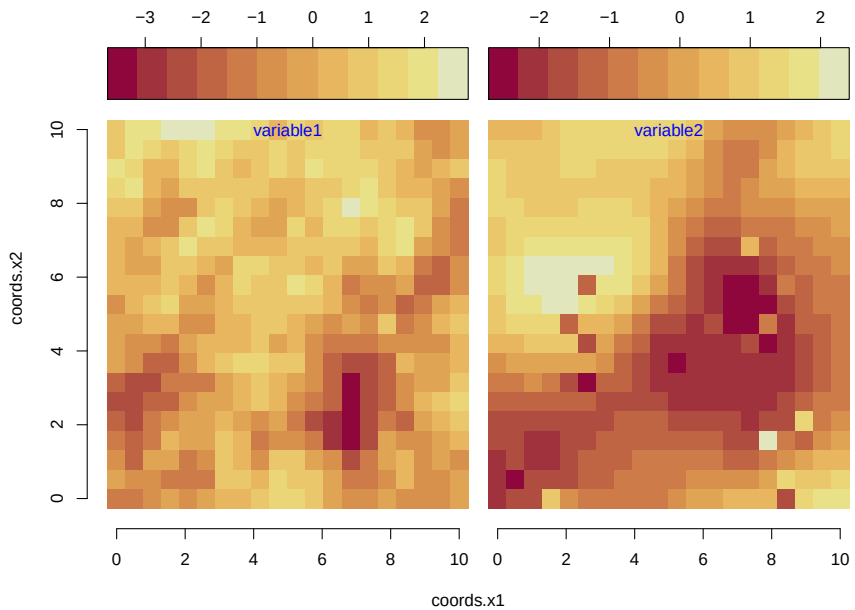
Simulations

Generation of spatial data of p variables for n locations (grid or Wallonia)

Simulation set-up

- Matérn model to generate spatial data
- Contamination by swapping observations with high/small PCA scores[3]





[3] Harris *et al.* (2014)

Results

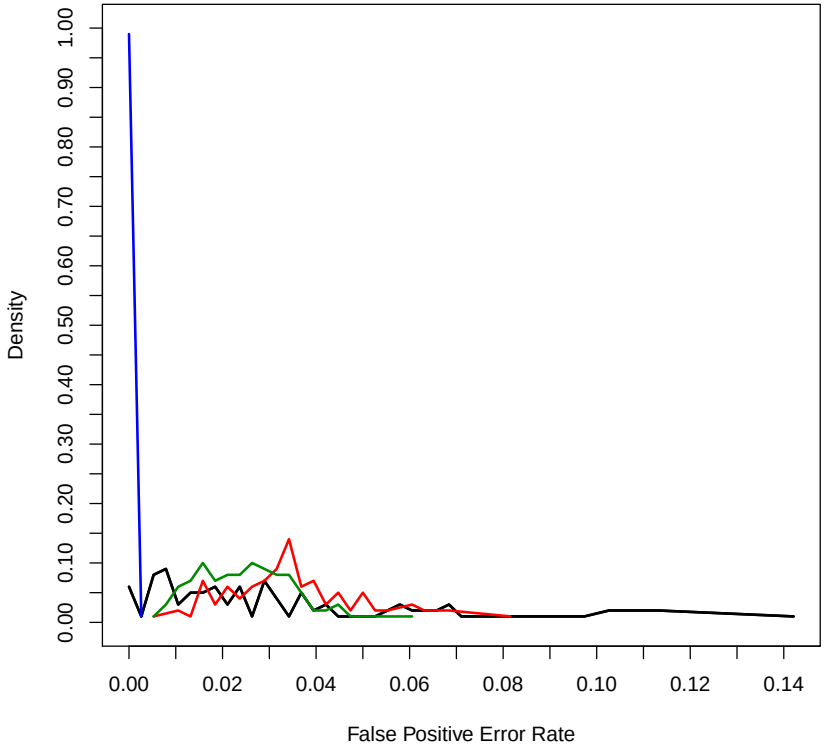
Performance criteria

- False Positive (FP): regular observations classified as local outliers.
- False Negative (FN): local outliers not detected.

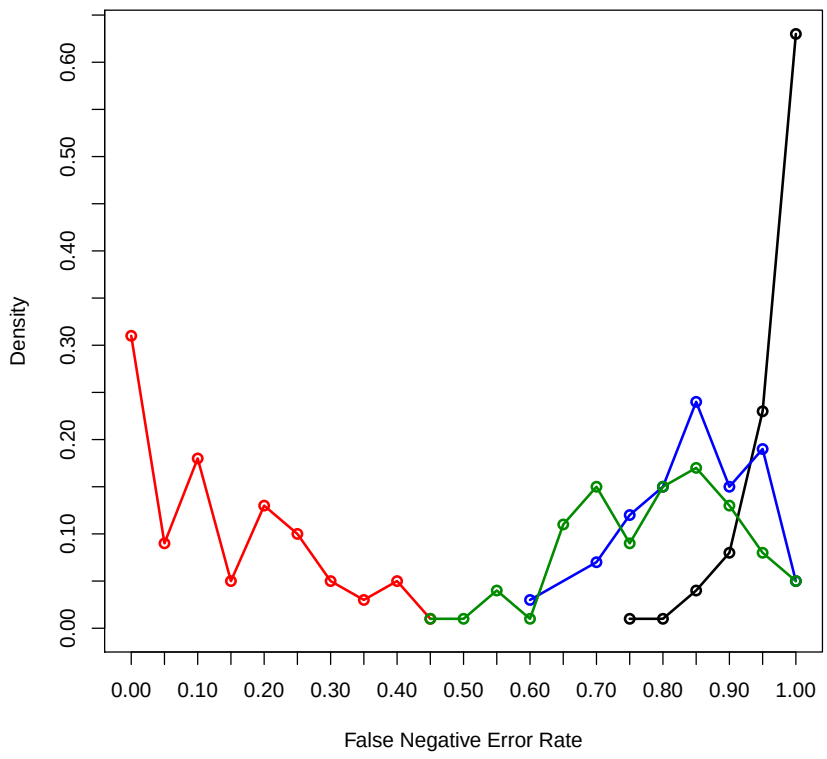
Goal: minimization of FP and FN
Priority: minimization of FP

Bivariate simulations on a grid

False Positive



False Negative

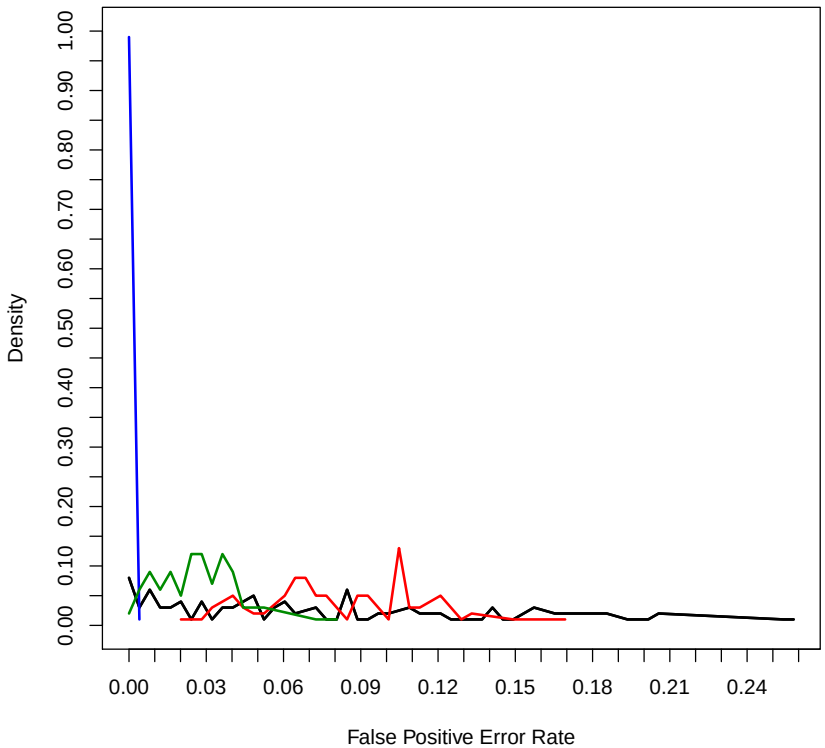


Chen *et al.* (2008) Filzmoser *et al.* (2014)

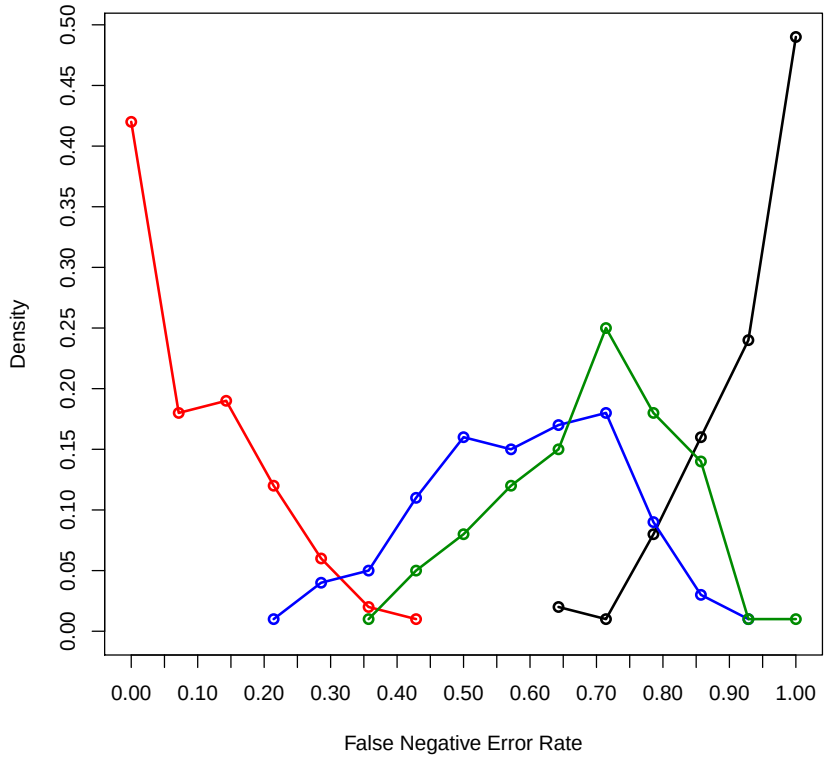
Harris *et al.* (2014) Regularized Filzmoser

Bivariate simulations on Walloon municipalities

False Positive



False Negative



Chen *et al.* (2008) Filzmoser *et al.* (2014)
Harris *et al.* (2014) Regularized Filzmoser

Ranking for these bivariate simulations

False Positive Error Rate

1. Filzmoser *et al.* ($< 5\%$)
2. Regularized Filzmoser ($< 5\%$)
3. Chen *et al.*
4. Harris *et al.*

False Negative Error Rate

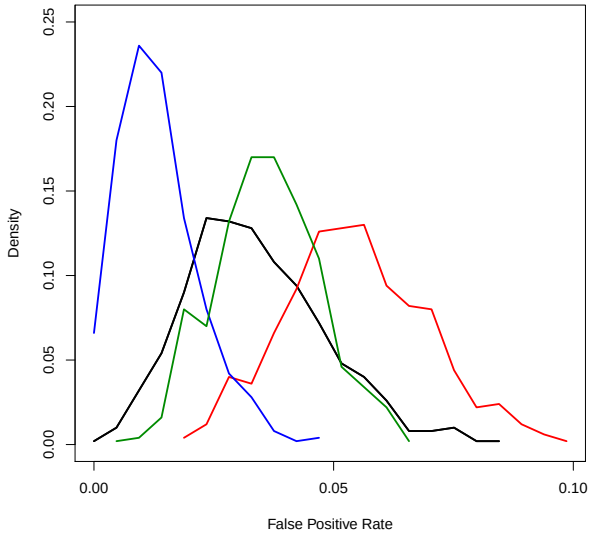
1. Chen *et al.*
2. Filzmoser *et al.* and its regularization
4. Harris *et al.*

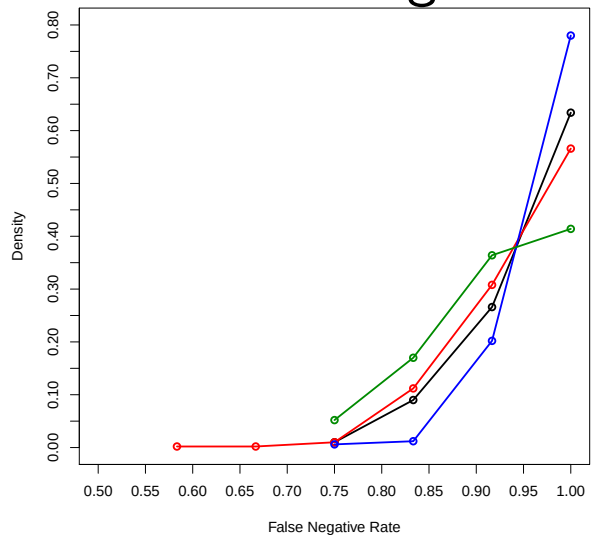
Simulations in 5 dimensions[4]

False Positive

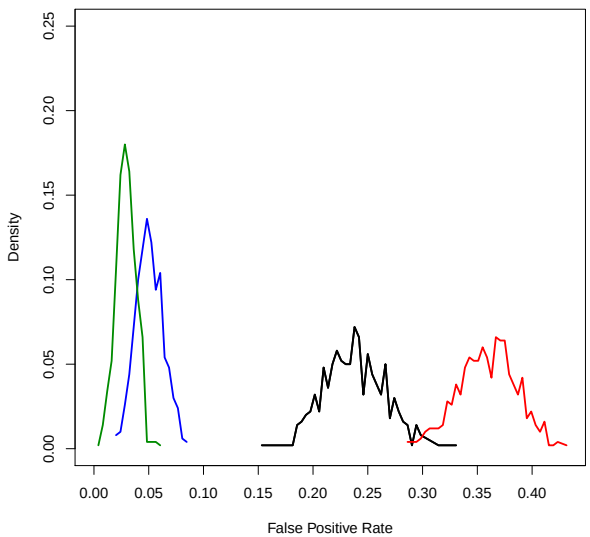
False Negative

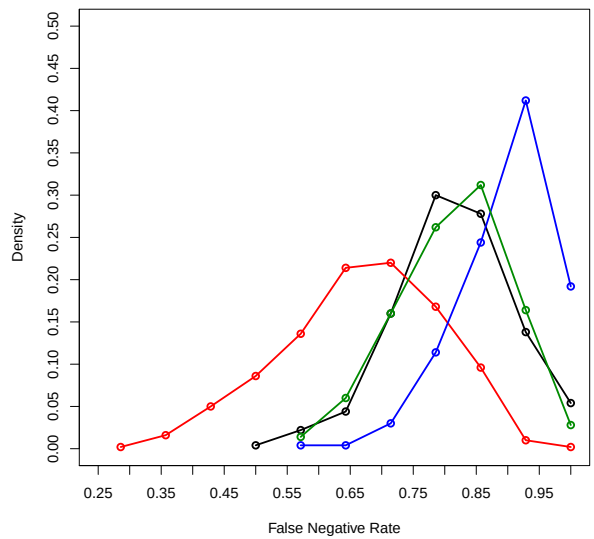
Grid





Wallonia





Chen *et al.*

Filzmoser *et al.*

Harris *et al.*

Regularized Filzmoser

[4] Parameters defined in Harris *et al.* (2014)

Conclusion of the preliminary simulations study

- Filzmoser *et al.* (2014) and its regularization perform better than the two other techniques, especially on irregular spatial domains.
- The regularization tends to increase the FP rate wrt the intial technique, but it gets better as the dimension increases.

Future work

The simulations study provides an objective way for comparing the detection techniques but...

- Other configurations need to be considered (higher dimensions, other correlation structures, other spatial set-ups,...).
- Other performance criteria might be useful to add to the FP and FN measures.
- The real-life application should be further explored to interpret in an economic way the local outliers detected.

References

- Chen, D. , Lu, C.T., Kou, Y. and Chen, F.,
On Detecting Spatial Outliers, *Geoinformatica* (2008).
- Filzmoser, P., Ruiz-Gazen, A. and Thomas-Agnan, C.,
Identification of Local Multivariate Outliers, *Statistical Papers* (2014).
- 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 Computing and Computer-Assisted Intervention-MICCAI* (2011).
- Harris, P., Brunsdon, C., Charlton, M., Juggins, S., and Clarke,
A Multivariate Spatial Outlier Detection Using Robust Geographically Weighted Methods, *Mathematical Geosciences* (2014).
- Haslett, J., Brandley, R., Craig, P., Unwin, A. et Wills, G.,
Dynamic Graphics for Exploring Spatial Data with Applications to Locating Global and Local Anomalies, *The American Statistician* (1991).
- Schubert, E., Zimek, A., and Kriegel, H.-P.,
Local Outlier Detection Reconsidered: A Generalized View on Locality with Applications to Spatial, Video, and Network Outlier Detection, *Data Mining and Knowledge Discovery* (2014).