Robust detection of local outliers in multivariate spatial data

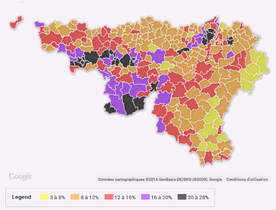
ERNST Marie Department of Mathematics, University of Liege, Belgium m.ernst@ulg.ac.be

Joint work with HAESBROECK Gentiane

Spatial data

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

Example: Unemployment rate in Walloon municipalities.



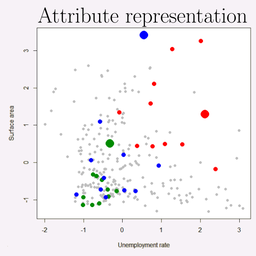
Local Outliers

In spatial data, Haslett et al. (1991) distinguishes two types of outliers.

- A **global outlier** has non spatial attributes with significantly differing values w.r.t. the majority of the data set $\{z_1, \ldots, z_n\}$.
- An observation at location s_i is a **local outlier** if its non spatial attributes z_i significantly lie far from the values z_j taken by its neighbors $(s_j \in \mathcal{N}_i)$.

Geographic representation

Attribute representation



- Hannut is a local but not global outlier.
- Tournai is a local and global outlier.
- Viroinval is a global but not local outlier.

Review of local and multivariate detections

Focus here on local outliers detection since global outliers detection does not consider geographical components and so usual outliers detection techniques can be used.

1. Median Algorithm, Chen et al. (2008)

Context	Model function
$\mathcal{N}_i$	Computation of h_i which is the difference (in $\mathbb{R}^p$) between z_i and the vector of marginal medians computed on z_j with $s_j \in \mathcal{N}_i$.
Reference	**Comparison function**
Global	Computation of the distances $d_{\hat{\mu},\widehat{\Sigma}}(h_i), 1 \leq i \leq n$, where $\hat{\mu}$ and $\widehat{\Sigma}$ are the MCD location and dispersion estimators computed on $\{h_1, \ldots, h_n\}$. Comparison of these distances with a F-quantile.

2. Detection technique of Filzmoser et al. (2014)

Context	Model function
Global	Robust estimation of the center and dispersion of $\{z_1, \ldots, z_n\}$ by means of the MCD estimator; yielding $\hat{\mu}$ and $\widehat{\Sigma}$.
$\mathcal{N}_i$	Computation of the n_i distances $d_{z_i,\widehat{\Sigma}}(z_j)$ with $s_j \in \mathcal{N}_i$. Computation of the isolation degree of s_i.
Reference	**Comparison function**
Global	Comparison of the isolation degrees and selection of the largest ones.

3. Geographically weighted detection, Harris et al. (2014)

Context	Model function
Global (optional)	Reduction of the dimension with robust PCA.
$\mathcal{N}_i$	Application of a Geographically Weighted PCA in $\mathcal{N}_i$. Computation of score distances (SD), orthogonal distances (OS) and component scores (CS).
Reference	**Comparison function**
Global	Comparison of the univariate measures SD, OS, and CS with theoretical/empirical quantiles.

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 a specific subset H of $\{1, \ldots, n\}$ that minimizes 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, L^1 or L^2 norm). The covariance matrix estimator is invertible but not robust.

- **Regularized MCD (Fritsch et al. (2011))**

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

for the optimal subset H.

New detection technique for local outliers

A two-step improvement is proposed to regularize the detection of Filzmoser et al. (2014).

1. Local structure

The n_i pairwise distances in $\mathcal{N}_i$ rely on the robust estimation of the global correlation structure but may be inefficient for neighborhoods of different shapes.

Improvement: using local estimation $\widehat{\Sigma}_i$ of the covariance matrix based on the regularized MCD. Indeed, as the size n_i may be smaller than p, regularization is needed for positive-definiteness of $\widehat{\Sigma}_i$ and the detection of outliers requires robustness.

2. Restriction to homogeneous neighborhoods

An observation should not be classified as a local outlier if its non spatial attributes differ from those of its neighbors because they are simply lying in an unstable area. Therefore one needs to take into account the possible heterogeneity of the neighborhoods whatever their size n_i.

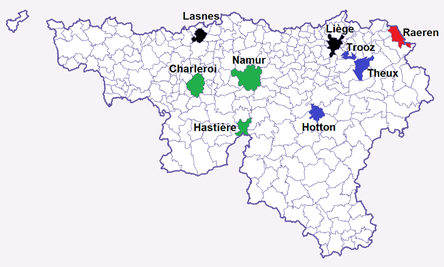
Improvement: let consider $\mathcal{N}_i \cup \{s_i\}$ and define

$$c_i = \text{mean}_{j \in H_i} \{ d_{\hat{\mu}_i, \widehat{S}_i}(z_j) \}$$

where $(\hat{\mu}_i, \widehat{\Sigma}_i)$ are estimated with regularized MCD, H_i is the corresponding optimal subset and $\widehat{S}_i$ is proportional to $\widehat{\Sigma}_i$ such that $\det(\widehat{S}_i) = 1$.

Only the spatial units having neighborhoods characterized by a c_i ranked among the $\beta \times n$ smallest are selected for the final step of the detection.

Comparison on the example



Dataset: 14 economic and demographic variables on the 262 Walloon municipalities.

Outliers detected with the different techniques:

Chen et al. (2008) Harris et al. (2014)

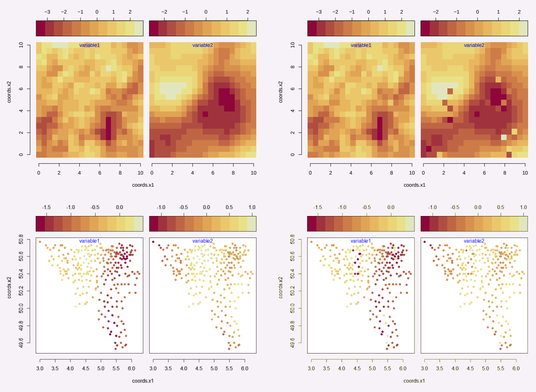
Filzmoser et al. (2014) Regularization

Simulations

In order to compare in an objective way, spatial data of p variables for n locations (grid or Wallonia) are simulated.

Simulation set-up: (Harris et al. (2014))

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



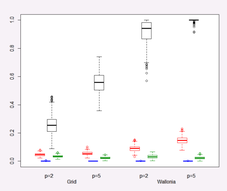
Performance criteria:

False Positive (FP): uncontaminated observations classified as local outliers.

False Negative (FN): contaminated observations not detected.

According to the simulation settings, the goal is to minimize the false positive error rate as a priority.

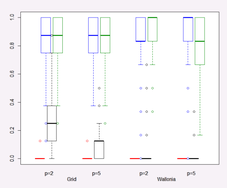
False Positive Error Rate



Preliminary results

Filzmoser et al. (2014) and its regularization perform better than the two other techniques, especially on irregular spatial domains.

False Negative Error Rate



The regularization tends to increase the FP rate w.r.t. the initial technique, but it gets better as the dimension increases. Their FN rate are similar. However one can expect a slight enhancement for the regularization on the Walloon municipalities with larger dimension.

On going research

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, ...). Secondly the real-life application should be further explored to interpret in an economic way the local outliers detected. A perspective for future work is to properly analyze the robust properties of the regularized MCD estimator and the distribution of the corresponding Mahalanobis distances.