Identification of local multivariate outliers

Anne Ruiz-Gazen and Christine Thomas-Agnan

Gremaq, TSE and IMT
Toulouse, France

(in collab. with Peter Filzmoser)

SSIAB - Avignon - 11/05/12

Introduction

In robust statistics, an observation is considered as outlying if it differs from the main bulk of the data set.

Introduction

In robust statistics, an observation is considered as outlying if it differs from the main bulk of the data set.

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon G$$

In the case of continuous attributes, the main bulk of the data set assumed to follow an elliptical distribution (e.g. gaussian) F and the outlying observations following a distribution G (e.g. point mass).

Introduction

In robust statistics, an observation is considered as outlying if it differs from the main bulk of the data set.

$$F_\varepsilon = (1-\varepsilon)F + \varepsilon G$$

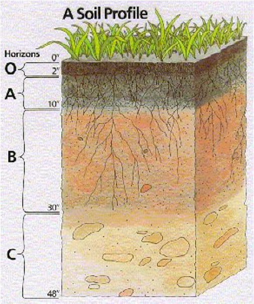
In the case of continuous attributes, the main bulk of the data set assumed to follow an elliptical distribution (e.g. gaussian) F and the outlying observations following a distribution G (e.g. point mass).

Objective : identify/detect

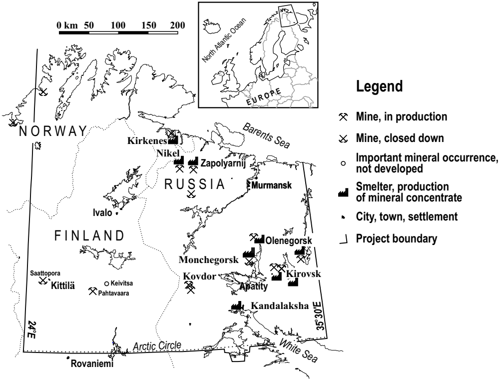
- gross errors,
- atypical observations

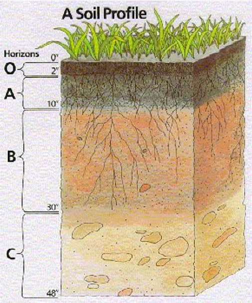
taking into account the multivariate and the spatial nature of the data.

Introduction



The Kola project : concentration measures for more than 50 chemical elements in four layers and 617 observations.





The Kola project : concentration measures for more than 50 chemical elements in four layers and 617 observations.

Data available in the R-package mvoutlier by M. Gschwandtner et P. Filzmoser.

1. Detection of outliers in a non spatial context
   - Detection of univariate outliers
   - Detection of multivariate outliers

2. Spatial outliers
   - Global and local outliers
   - Identification of univariate spatial outliers

3. Identification of multivariate spatial outliers
   - Variocloud of pairwise Mahalanobis distances
   - Toy example
   - Quantile geographical-variate plot

Detection of univariate outliers

Let us consider a data set x, $n \times p$ with n observations x_i and p variables.

Detection of univariate outliers

Let us consider a data set x, $n \times p$ with n observations x_i and p variables. In one dimension ($p = 1$), the detection of outliers is often based on

$$\frac{|x_i - \bar{x}|}{\sigma_x}$$

(Grubbs, 1969).

Detection of univariate outliers

Let us consider a data set x, $n \times p$ with n observations x_i and p variables. In one dimension ($p = 1$), the detection of outliers is often based on

$$\frac{|x_i - \bar{x}|}{\sigma_x}$$

(Grubbs, 1969).

Problem of masking effect : outliers may spoil the empirical mean and the standard deviation estimators in such a way that outliers are not detected.

Detection of univariate outliers

Let us consider a data set x, $n \times p$ with n observations x_i and p variables. In one dimension ($p = 1$), the detection of outliers is often based on

$$\frac{|x_i - \bar{x}|}{\sigma_x}$$

(Grubbs, 1969).

Problem of masking effect : outliers may spoil the empirical mean and the standard deviation estimators in such a way that outliers are not detected.

Robust version : $\bar{x}$ and σ_x replaced by some robust estimators such as the median and the MAD.

Detection of univariate outliers

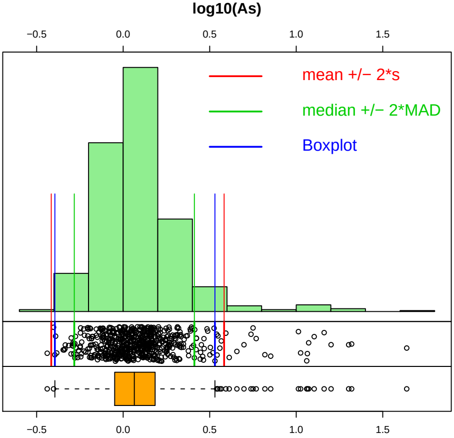


FIG.: Histogram of log(Arsenic)

Detection of multivariate outliers

In the multivariate case ($p > 1$), detection based on Mahalanobis distances to the center of the data set :

Detection of multivariate outliers

In the multivariate case ($p > 1$), detection based on Mahalanobis distances to the center of the data set :

Let t be a location estimator ($p \times 1$) and C a dispersion matrix estimator ($p \times p$) of the distribution of the main bulk of the data.

$$\mathrm{MD}(x_i, t, C) = \left\{(x_i - t)' C^{-1} (x_i - t)\right\}^{1/2}$$

Detection of multivariate outliers

In the multivariate case ($p > 1$), detection based on Mahalanobis distances to the center of the data set :

Let t be a location estimator ($p \times 1$) and C a dispersion matrix estimator ($p \times p$) of the distribution of the main bulk of the data.

$$\mathrm{MD}(x_i, t, C) = \left\{(x_i - t)' C^{-1} (x_i - t)\right\}^{1/2}$$

Multivariate outliers are associated with large values of Mahalanobis distances.

Detection of multivariate outliers

In the multivariate case ($p > 1$), detection based on Mahalanobis distances to the center of the data set :

Let t be a location estimator ($p \times 1$) and C a dispersion matrix estimator ($p \times p$) of the distribution of the main bulk of the data.

$$\mathrm{MD}(x_i, t, C) = \left\{(x_i - t)' C^{-1} (x_i - t)\right\}^{1/2}$$

Multivariate outliers are associated with large values of Mahalanobis distances.

In the Gaussian case $\mathcal{N}(\mu, \Sigma)$, the $\mathrm{MD}^2(x_i, \mu, \Sigma)$ follow a chi-square distribution with p degrees of freedom and a common used cut-off value is the quantile of order 95% of this chi-square distribution.

Detection of multivariate outliers

(Rousseeuw and Van Zomeren, 1990)

Detection of multivariate outliers

(Rousseeuw and Van Zomeren, 1990)

Use robust estimators t and C such as the Minimum Covariance Determinant (MCD) estimators. Look for a subset of data points (e.g. 75%) having the smallest determinant for its covariance matrix.

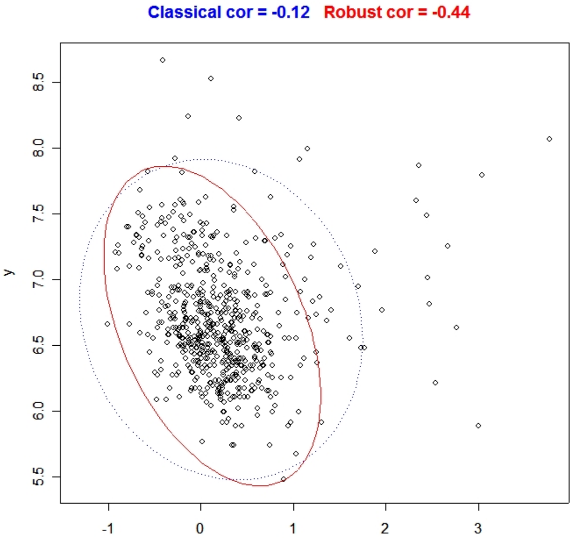
Detection of multivariate outliers

(Rousseeuw and Van Zomeren, 1990)

Use robust estimators t and C such as the Minimum Covariance Determinant (MCD) estimators. Look for a subset of data points (e.g. 75%) having the smallest determinant for its covariance matrix.

R-packages mvoutliers et rrcov.

In two dimensions, scatterplot with ellipsoids corresponding to non-robust estimators (blue) t and C and MCD estimators (red) for a quantile of order 95%.



Global and local outliers

Global outlier : relative to the distribution of the whole data set (whole area of study).

Global and local outliers

Global outlier : relative to the distribution of the whole data set (whole area of study).

Local outlier : relative to the sub-distribution associated with the observation and its neighborhood.
Underlying assumption : positive spatial autocorrelation.

Exploratory plots for identifying univariate spatial outliers

- Neighbor plot
- Moran plot
- Drift map
- Angle plot
- Variocloud

Exploratory plots for identifying univariate spatial outliers

- Neighbor plot
- Moran plot
- Drift map
- Angle plot
- Variocloud

R-package GeoXp by C. Thomas et al. (2012).

Variocloud of pairwise Mahalanobis distances

$$\mathrm{MD}(x_i, x_j, C) = \left\{(x_i - x_j)' C^{-1} (x_i - x_j)\right\}^{1/2}$$

Variocloud of pairwise Mahalanobis distances

$$\mathrm{MD}(x_i, x_j, C) = \left\{(x_i - x_j)' C^{-1} (x_i - x_j)\right\}^{1/2}$$

Draw a variocloud by replacing absolute pairwise difference with robust pairwise Mahalanobis distances (MCD covariance estimator)

Variocloud of pairwise Mahalanobis distances

$$\mathrm{MD}(x_i, x_j, C) = \left\{(x_i - x_j)' C^{-1} (x_i - x_j)\right\}^{1/2}$$

Draw a variocloud by replacing absolute pairwise difference with robust pairwise Mahalanobis distances (MCD covariance estimator)

Draw only part of the cloud, summarize the rest by conditional quantile curves

Example of multivariate variocloud

Selected units : small spatial distances and high pairwise Mahalanobis distance

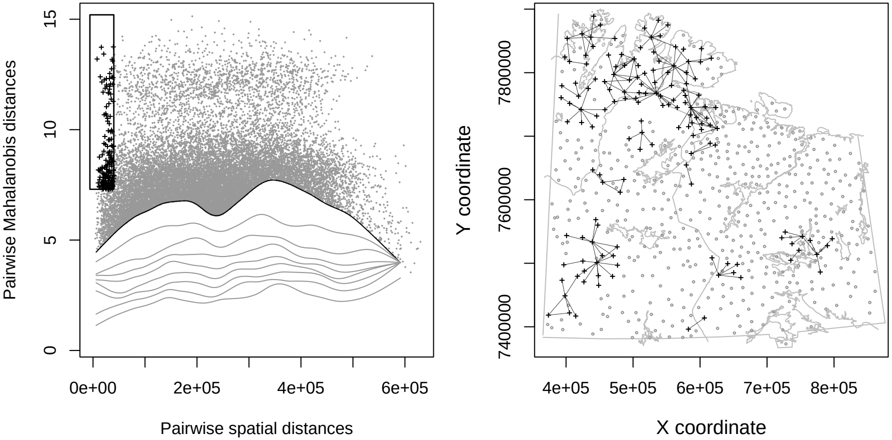


Fig. : Multivariate variocloud

A small toy example

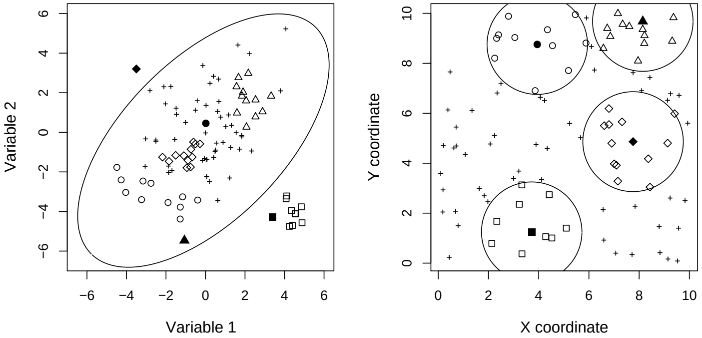


FIG.: Toy example

Comparing pairwise geographical distance and pairwise distance in the non spatial attributes space.

Variocloud for the toy example

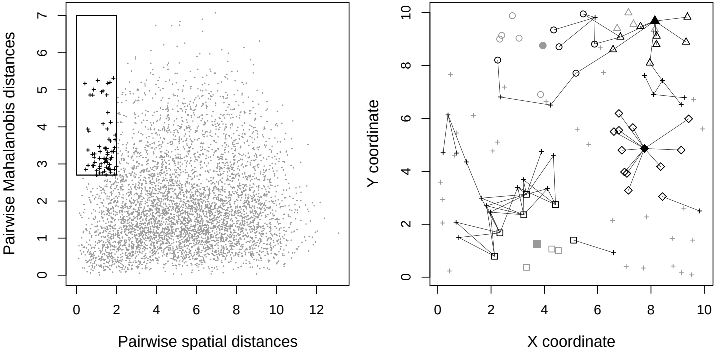


FIG.: Toy example

Distribution of the pairwise Mahalanobis Distance

When the observations $X_1, \ldots, X_n$ are i.i.d. with a normal distribution $\mathcal{N}(\mu, \Sigma)$, we can prove that :

Conditional on one observation X_i, the distribution of the squared pairwise Mahalanobis distance

$\mathrm{MD}^2(X_i, X_j, \Sigma)$ of X_j, $j \neq i$ is a non central chi-square distribution

with the Mahalanobis distance $\mathrm{MD}^2(X_i, \mu, \Sigma)$ as the non-centrality parameter and p degrees of freedom.

Distribution of pairwise MD : toy example

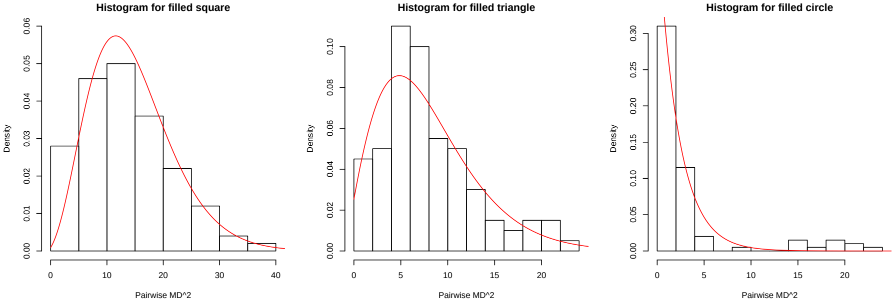


FIG.: Toy example

Quantile geographical-variate plot

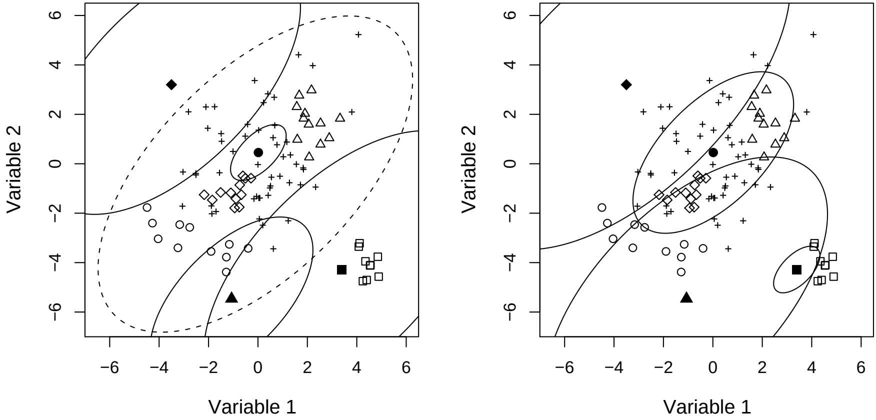


FIG.: Toy example

Quantile geographical-variate plot

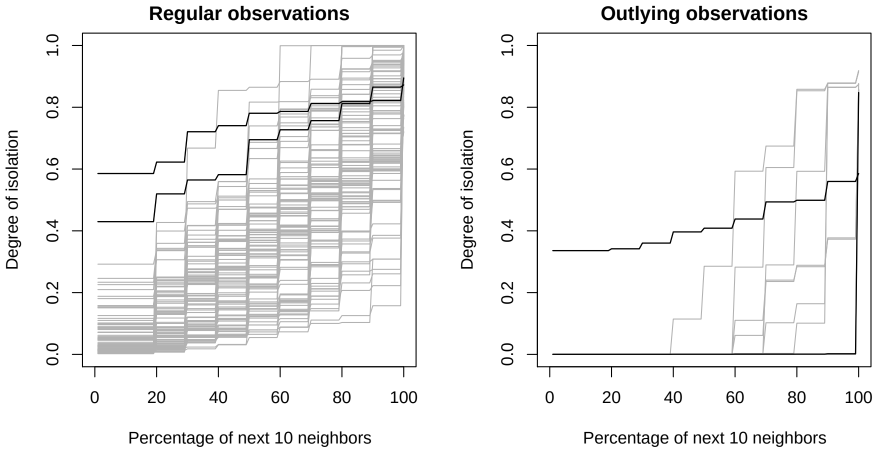


FIG.: Toy example

Quantile geographical-variate plot on Kola example

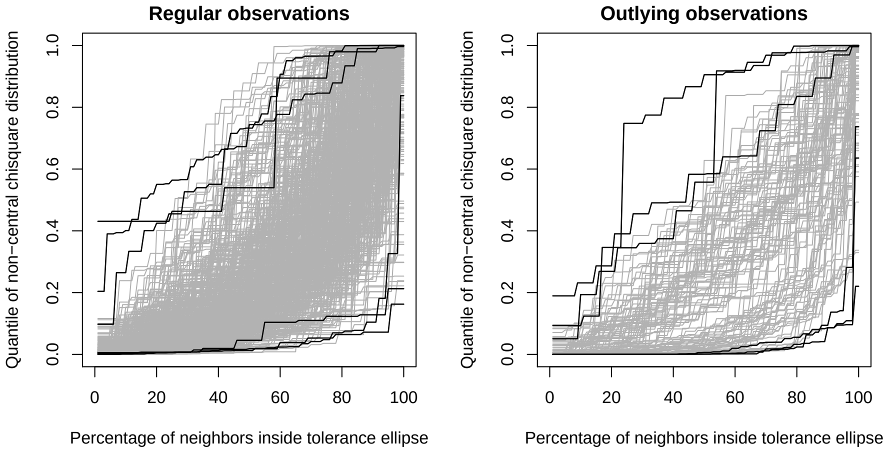


FIG.: Kola example

Conclusion

Thank you for your attention !