Assessing the Impact of Outliers on Least Square Variogram Model

Caterina Daidone, Marco Zanotti

University Milano-Bicocca

Contents

1. Introduction

Geostatistics

Gaussian Random Field

2. Methods

The **classical variogram** estimator based on Matheron (1962) is defined as:

$$2\hat{\gamma} = \frac{1}{|N(h)|} \sum_{N(h)} \left(Z(s_i) - Z(s_j)\right)^2, \quad h \in \mathbb{R}^d$$

where $N(h) = \{(s_i, s_j) : s_i - s_j = h; i, j = 1, ..., n\}$ and $|N(h)|$ is the number of distinct pairs in $N(h)$.

The **robust variogram** estimator based on Cressie (1980) is defined as:

$$2\bar{\gamma}(h) \equiv \left\{ \frac{1}{|N(h)|} \sum_{N(h)} |Z(s_i) - Z(s_j)|^{1/2} \right\}^4 /(0.457 + 0.494/|N(h)|)$$

and

$$2\tilde{\gamma}(h) = [med\{|Z(s_i) - Z(s_j)|^{1/2} : (s_i, s_j) \in N(h)\}]^4 / B(h)$$

where $med\{\cdot\}$ is the median of the sequence and $B(h)$ corrects for bias (asymptotically 0.457).

3. Simulation & Results

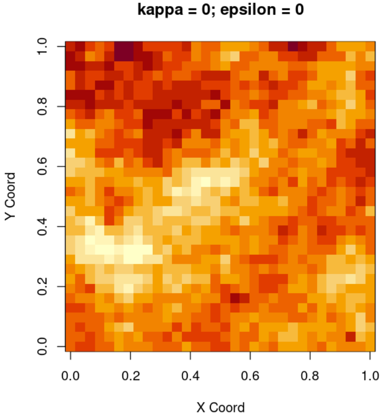
Following Hawkins (1984), the departure from Gaussianity is obtained simulating a **GRF** with probability $1-\epsilon$ and a **CGRF** with probability ϵ.

$$\begin{cases} Gau(0, c_o), & with\ probability\ 1-\epsilon \\ Gau(0, k^2 c_0), & with\ probability\ \epsilon \end{cases}$$

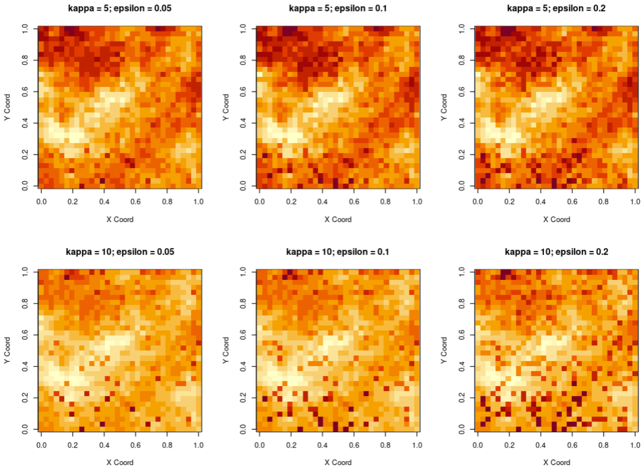
where ϵ is the probability of contamination and k measures the scale of the contamination.

To practically simulate the underlying GRF, the **grf** function of the **geoR** package in R is used.

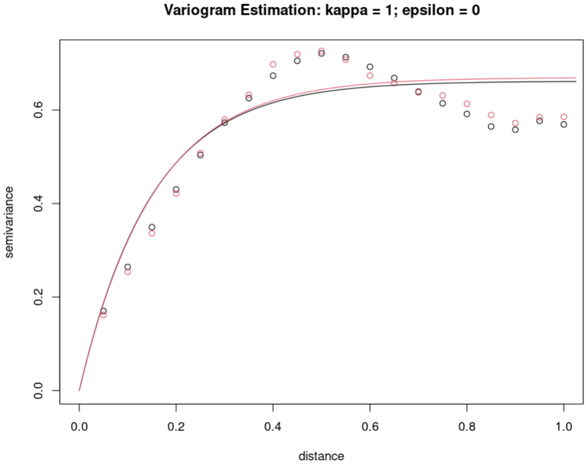
The **base scenario** represents no contamination and is simulated with $\epsilon = 0$.



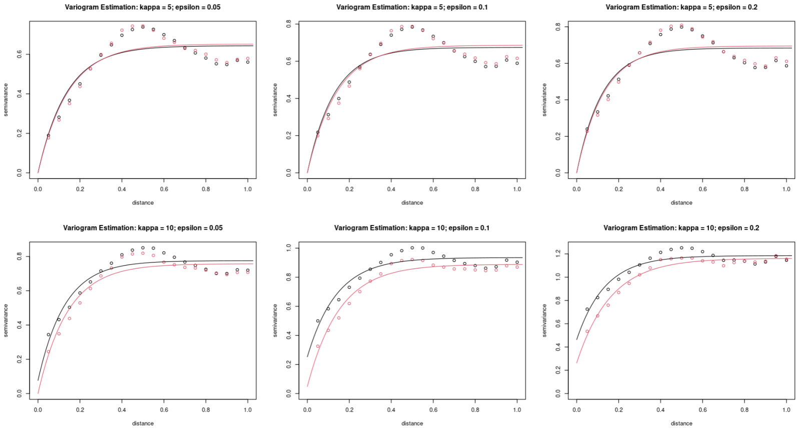
Six different contaminated scenarios based on the combinations of $epsilon = (0.05, 0.1, 0.2)$ and $kappa = (5, 10)$ are simulated.



In the **base scenario** the two methods provide almost **identical results**.



In the presence of **contamination** the two methods provide **different results**.



4. Conclusions

The theoretical considerations suggest that the robust variogram is less sensitive to the presence of outliers. For this reason it should be preferred when the data are contaminated.

The simulation study confirms this results and shows that:

- the **robust variogram** yields more **stable estimates** when the **scale** of the contamination **increases**
- if the **scale** of the contamination is **small**, the two methods provide **similar results**.

Bibliografy

Cressie N. 1993. Statistics for Spatial Data. John Wiley and Sons Inc., New York

Thank you!