Assessing the Impact of Outliers on Least Square Variogram Model

Caterina Daidone, Marco Zanotti

University Milano-Bicocca

Contents

1. Introduction

Geostatistics

Spatial data are observations measured at known locations, based on the notion that near things are more related than distant things.

The types of spatial data are areal data, point pattern, and geostatistical data.

A Geostatistical model may be represented in the form

$$Y(x) = \mu(x) + S(x) + W(x), \quad x \in D$$

where $\mu(x) =$ trend (large scale component), $S(x) =$ 0-mean spatial stochastic process (small scale term), $W(x) =$ 0-mean n independent random variables with $Var(W(x)) = \tau^2$ (white noise) and D space index.

Gaussian Random Field

An infinite indexed family of random variables defined on a common probabilistic space. "Spatial" is deployed when the trajectory of the process is a deterministic function with 2-D or 3-D domain.

Three main ingredients: a index space, a probability space and a state space.

A stochastic process is Gaussian when follows a k-dimensional (multivariate) Normal distribution.

Main advantage:

- strong stationarity and second-order stationarity coincide (the invariance of the distributional features of the process).

2. Variogram Methods

The variogram function $2\gamma(\cdot)$ (treated as a parameter of a stochastic process) verifies spatial dependence in geostatistics. It is defined as

$$2\gamma(s_1 - s_2) \equiv var(Z(s_1) - Z(s_2))$$

$2\gamma(\cdot)$ is a function only of the increment $s_1 - s_2 = h$ and it will be treated as a parameter of a stochastic process, restricted to be symmetric about 0 and conditionally-negative definite.

Three basic models in terms of semivariogram ($\gamma(\cdot)$):

- linear: $\gamma(h;\theta) = c_0 + b_l||h||$
- spherical: $\gamma(h;\theta) = c_0 + c_s\{(3/2)(||h||/a_s) - 1/2(||h||/a_s)^3$
- exponential: $c_0 + c_e\{1 - exp(-||h||/a_e)\}$

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

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

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 is based on Cressie (1980)

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

where a bias correction is used (asymptotically 0.457).

3. Simulation & Results

Following Hawkins (1984), the departure from Gaussianity is obtained simulating

$$Z(s) = \mu + W(s) + E(s)$$

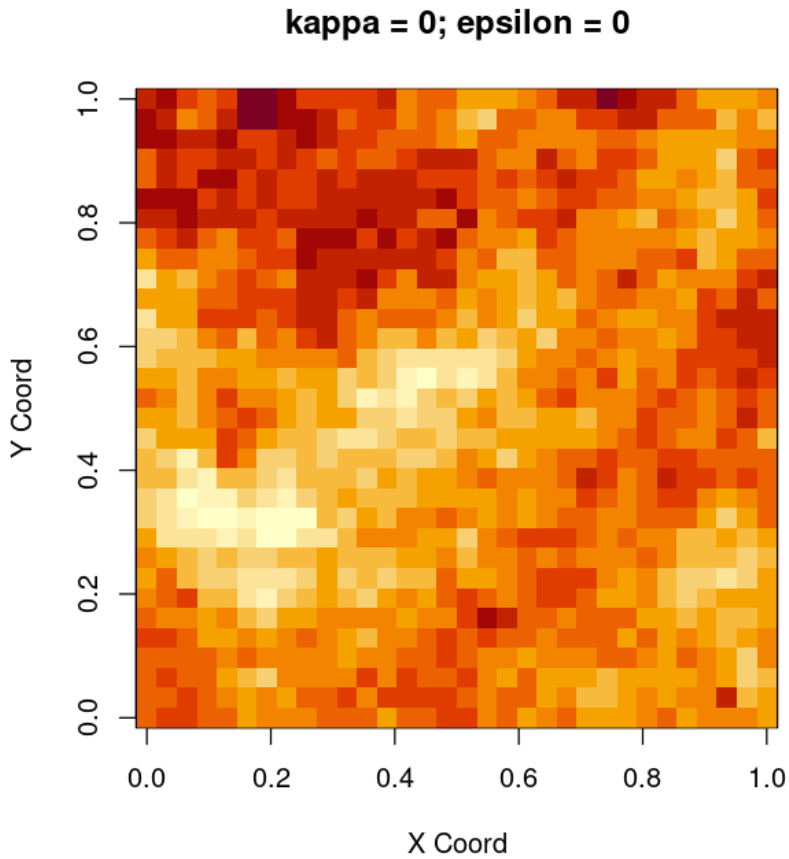
and $E(s)$ is a Gaussian process with probability $1 - \epsilon$ and a Contaminated Gaussian process with probability ϵ.

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

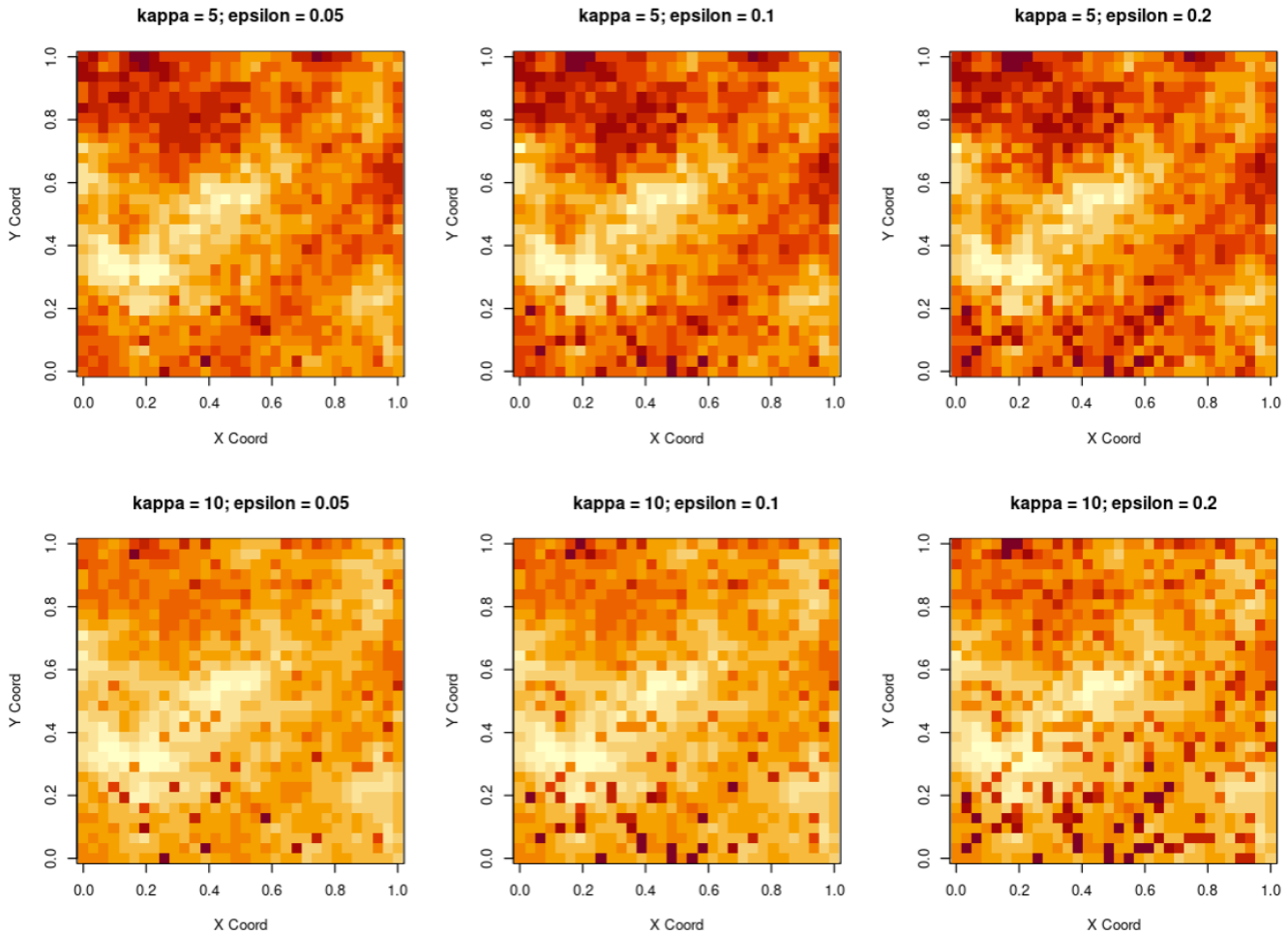
where ϵ is the probability of contamination and k measures the scale of the contamination.

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

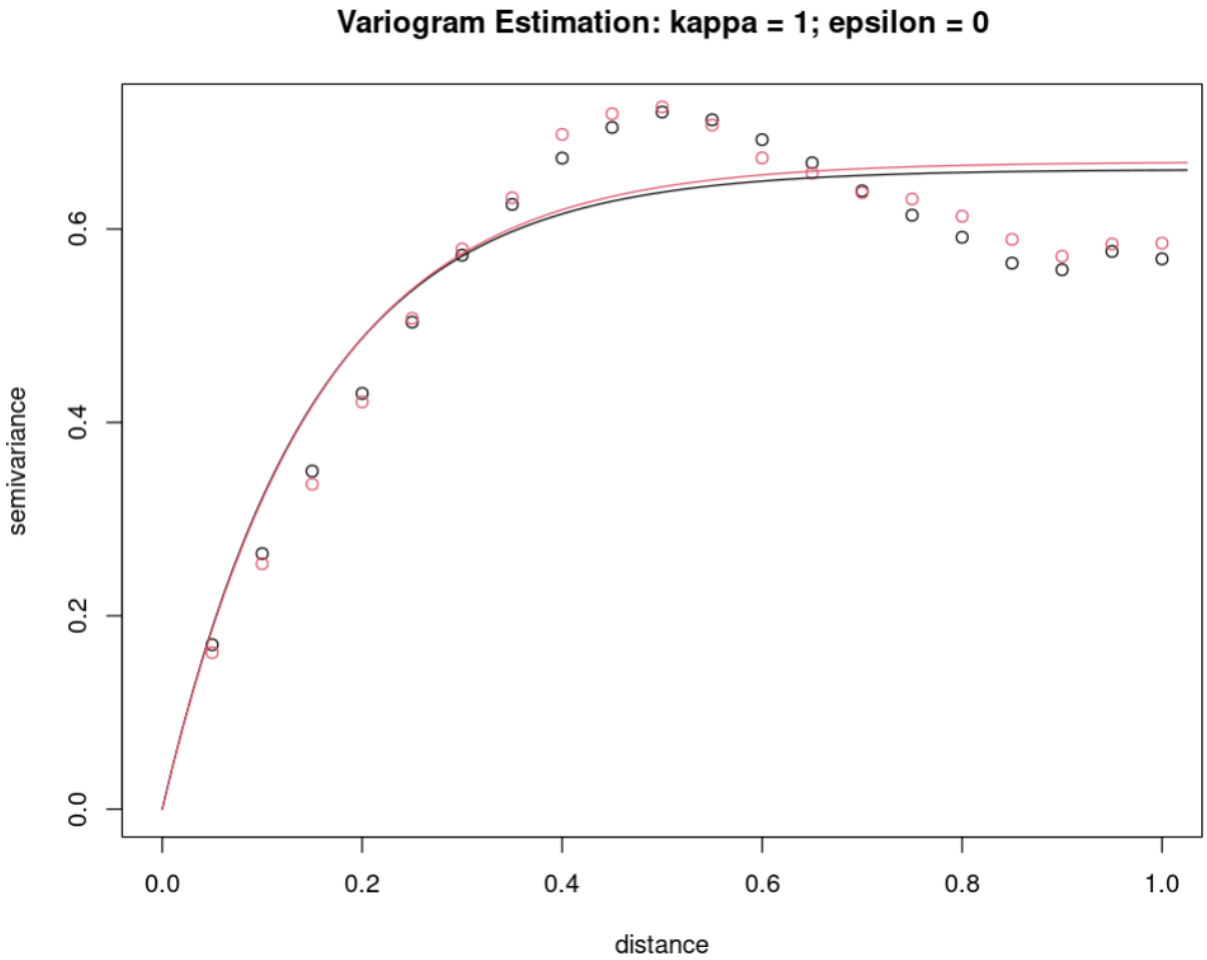
The **base scenario** represents no contamination and is simulated with $\epsilon = 0$.



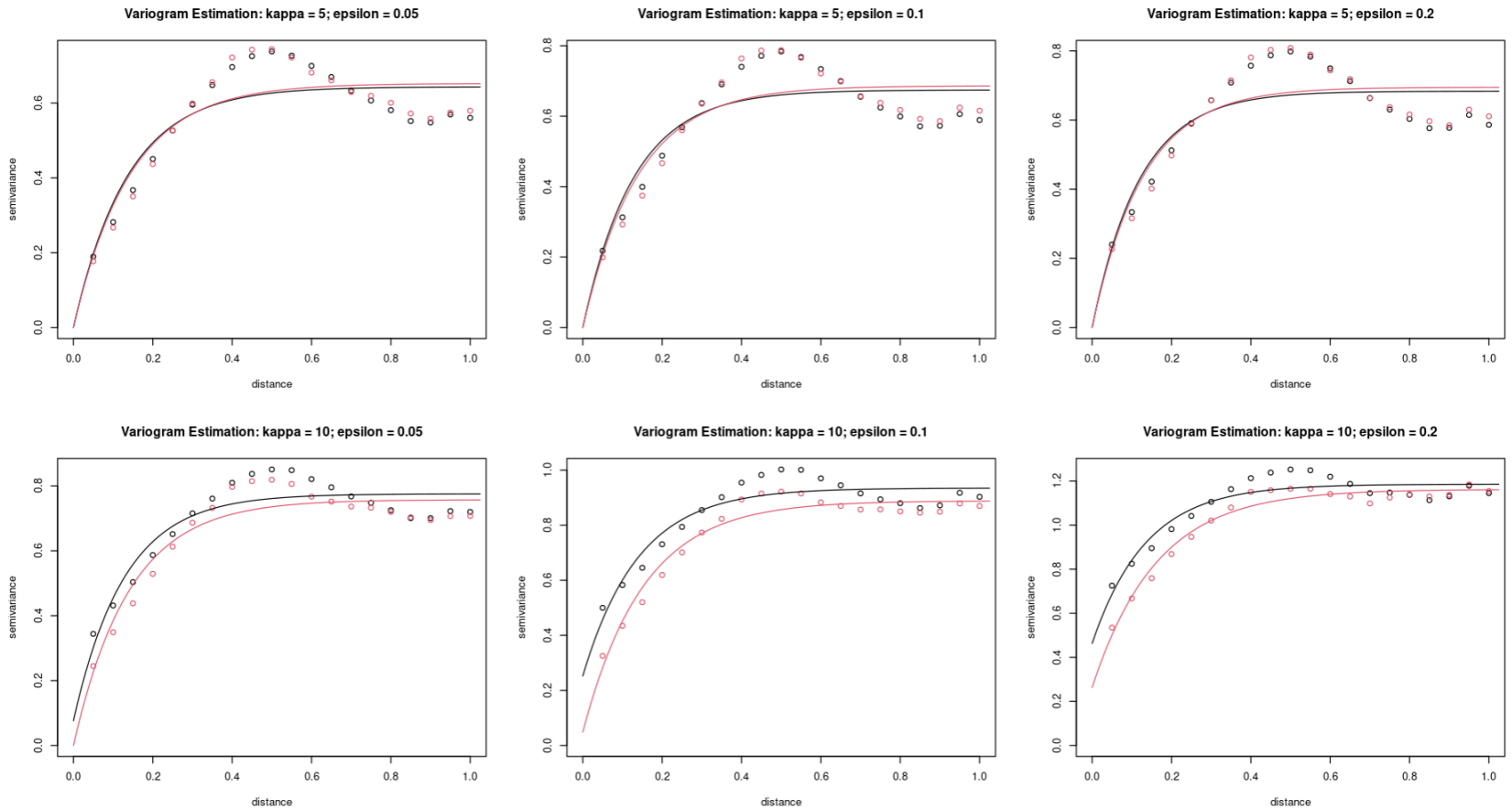
Six different contaminated scenarios based on the combinations of $epsilon = (0.05, 0.1, 0.2)$ and $kappa = (5, 10)$ are simulated.



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. & Hawkins, D. M. 1980, Robust estimation of the variogram: I, Journal of the international Association for Mathematical Geology 12, 115-125

Hawkins, D. M. & Cressie, N. 1984, Robust kriging—a proposal', Journal of the International Association for Mathematical Geology 16, 3-18

Matheron, G. 1962, Precision of exploring a stratified formation by boreholes with rigid spacing - application to a bauxite deposit, Elsevier, pp. 407-422

Thank you!