Assessing the Impact of Outliers on Least Square Variogram Model

Caterina Daidone, Marco Zanotti
University of Milano Bicocca

February 28, 2024

Abstract

It is a matter of common experience that ore values often do not follow the normal (or lognormal) distributions assumed for them, but, instead, follow some other heavier-tailed distribution. This study reviews the two most popular methods for the variogram estimation, that is the classical and the robust variograms. Moreover, a simulation study is performed to assess the impact of outliers on the variogram estimates. It is shown that the use of the robust variogram yields stable estimates when the scale of the contamination increases.

Keywords: geostatistics, outliers, variogram, robust statistics

1 Introduction

Spatial data consist of observations measured at known specific locations or within specific regions and show regularities in space. Remembering that the first law of geography states that "everything is related to everything else, but near things are more related than distant things" (Tobler (1969)), a statistical analysis should be considered to improve prediction and to adjust inference since the iid assumption is no more valid. Correlated data induce lost of efficiency (typically smaller standard errors) hence altering the significance of test procedures and the actual coverage of confidence intervals. Typical stochastic models for geostatistical data (a type of spatial data, among which areal data and point pattern data) are spatial stochastic processes, Gaussian when data are continuous. A statistical technique to obtain information from the stressed kind of spatial data is the geostatistical model, defined as

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

where $\mu(x)$ is the trend or drift (large scale component) of the process, $S(x)$ (small-scale term) is 0-mean spatial stochastic process with a given correlation structure, $W(x)$ is white noise, namely 0-mean independent random variables with $Var(W(x)) = \tau^2$, D is the space index. Moreover, observed variables often contain outliers that have unusually large or small values when compared with others and could affect the results of analysis in geostatistical data. Some methodology results about Gaussian random process and variogram estimation are presented. Additionally through a simulation study, a comparison between classical and robust methods is shown to detect some changes of the estimates of the variogram parameters, using least squares estimation.

2 Methods

2.1 Gaussian Random Field

A random field (RF) or stochastic random process is an infinite indexed family of random variables defined on a common probabilistic space (Ω, F, P). The word "spatial" is deployed when the trajectory of the process is a deterministic function with 2-D or 3-D domain. In this framework, the index is denoted with s. Three main ingredients are very important in RF: i) a parametric (or index) space D; ii) a probability space and iii) a state space, defined by the set of all possible values that each random variable at location s take on.

A stochastic process is Gaussian when for any set $x_1, ..., x_k$ in D and any integer $k \geq 1$ and K $\in \mathbb{N}, [S_1, ..., S_k]$ (the indexed family of random variables) follows a k-dimensional (multivariate) Normal distribution. A point that should be highlighted is that strong stationarity and second-order stationarity coincide in the Gaussian random field.
Strongly stationary is a very difficult to verify and restrictive assumption since founded on the equivalence of probability density functions between the "original process" and a translated process along the domain. For this reason the second-order or weakly stationarity condition, easier since based only on the moments of distribution, is analysed. In a wide sense, stationarity concerns the invariance of the distributional features of the process.

2.2 Variogram Estimation

The variogram function $2\gamma(\cdot)$ is an important quantity to verify spatial dependence in geostatistics. It is defined as

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

It is considerable to stress that $2\gamma(\cdot)$ is a function only of the increment $s_1 - s_2 = h$ and the variogram will be treated as a parameter of a stochastic process, restricted to be symmetric about 0 and conditionally-negative definite. $\gamma(\cdot)$ has been called as semivariogram by Matheron (1962).

When the process is intrinsically stationary, namely it satisfies the second-order stationarity

$(E(Z(s)) = \mu$, for all $s \in D)$ and the constant mean assumption $(var(Z(s_1) - Z(s_2)) = 2\gamma(s_1 - s_2)$, for all $s_1, s_2 \in D)$, the variogram may be defined also as $E(Z(s_1) - Z(s_2))^2$. Thus a weakly stationary is also intrinsic stationary, although the reverse does not hold true except for the Gaussian process. A RF with second-order stationary and intrinsic stationary is isotropic, otherwise anisotropic.

Three basic isotropic models in terms of semivariogram (Journel & Huijbregts (1978)) are considered:

- linear: $\gamma(h;\theta) = c_0 + b_l||h||$ when $h \neq 0$ and 0 otherwise $(\theta = (c_o, b_l)', c_0 \geq 0$ and $b_l \geq 0)$,
- spherical: $\gamma(h;\theta) = c_0 + c_s\{(3/2)(||h||/a_s) - 1/2(||h||/a_s)^3$ when $0 \leq ||h|| \leq a_s$, $c_0 + c_s$ when $||h|| \geq a_s$, 0 when $h = 0$ $(\theta = (c_0, c_s, a_s)'$, $c_0 \geq 0$, $c_s \geq 0$, $a_s \geq 0)$
- exponential: $c_0 + c_e\{1 - exp(-||h||/a_e)\}$, when $h \neq 0$, 0 otherwise $(\theta = (c_0, c_e, a_e)'$, $c_0 \geq 0$, $c_e \geq 0$, $a_e \geq 0)$.

The classical variogram estimator (Matheron (1962)), based on method of moments, is given by

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

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 $N(h)$.

2.3 Robust Variogram Estimation

The adjective robust refers to inference procedures stable also when model assumptions depart from those of central model, for instance a small contamination of a Gaussian random process. Cressie & Hawkins (1980) take fourth-roots of squared differences to yield robust estimators

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

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). $(Z(s_i) - Z(s_j))^2$ is a chi-squared random variable with one degree of freedom for Gaussian data. The power transformation that makes this most Gaussian-like is the square root of the absolute difference (the fourth root), namely $|(Z(s_i) - Z(s_j))|^{1/2}$. It is important to remark that the sums between classical and robust are not independent and when the dependence is higher, they are less efficient in estimating the variogram.

Robustness is a based-model concept and, following Hawkins & Cressie (1984) the model is given by

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

where $W(\cdot)$ is a zero-mean intrinsically stationary Gaussian process whose variogram is continuous at origin and $E(\cdot)$ is zero-mean white noise process, whose distribution is a contaminated Gaussian and the amount of contamination is given by a constant (ϵ). The bias of $\bar{\gamma}$ is less than $\hat{\gamma}$, although proportionally less when the uncontaminated part increases. A comparison plot of experimental variogram and fitted theoretical variogram is an invaluable diagnostic tool to measure the sum of squares of the differences between a generic variogram estimator ($2\gamma^*(he)$) and a model ($2\gamma(he;\theta)$) (Hawkins & Cressie (1984)).

The method of ordinary least squares specifies that θ is estimated by minimizing

$$\sum_{j=1}^{K} \{2\gamma^*(h(j)e) - 2\gamma(h(j)e;\theta)\}^2$$

for some direction e, where K are the number of lags.

3 Simulation

Simulating Gaussian Random Fields (GRF) and Contaminated Gaussian Random Fields (CGRF) involves different approaches due to the added complexity of contamination. GRFs solely need a mean and covariance function. This function dictates the smoothness and spatial dependence of the field. Common methods for generating GRF include the spectral method (using Fast Fourier Transforms) and the Cholesky decomposition. Simulating CGRF, instead, introduces an additional layer of complexity. In this case, the underlying GRF represents the "true" signal, while the contamination acts as an independent noise process. Common approaches involve generating the GRF first, then adding a separate noise field with its own properties (e.g., mean, variance, spatial dependence). Alternatively, one can directly simulate the CGRF by incorporating the contamination into the covariance function itself.

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

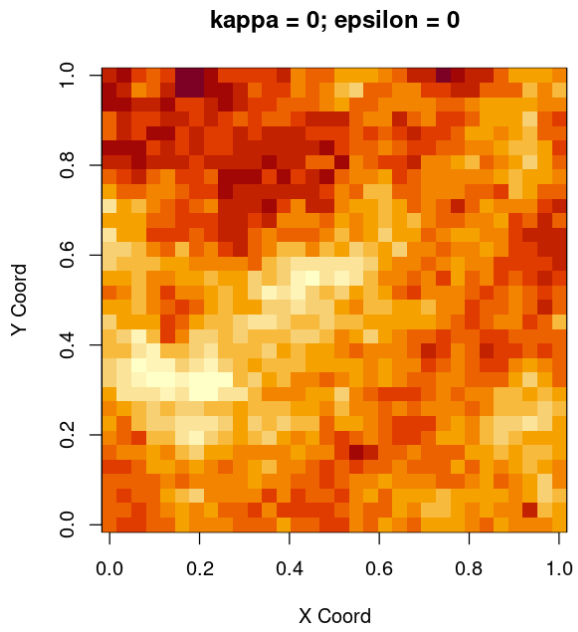
$$Z(s) = \mu + W(s) + E(s)$$

$E(s)$ is a GRF with probability $1 - \epsilon$ and a CGRF with probability ϵ.

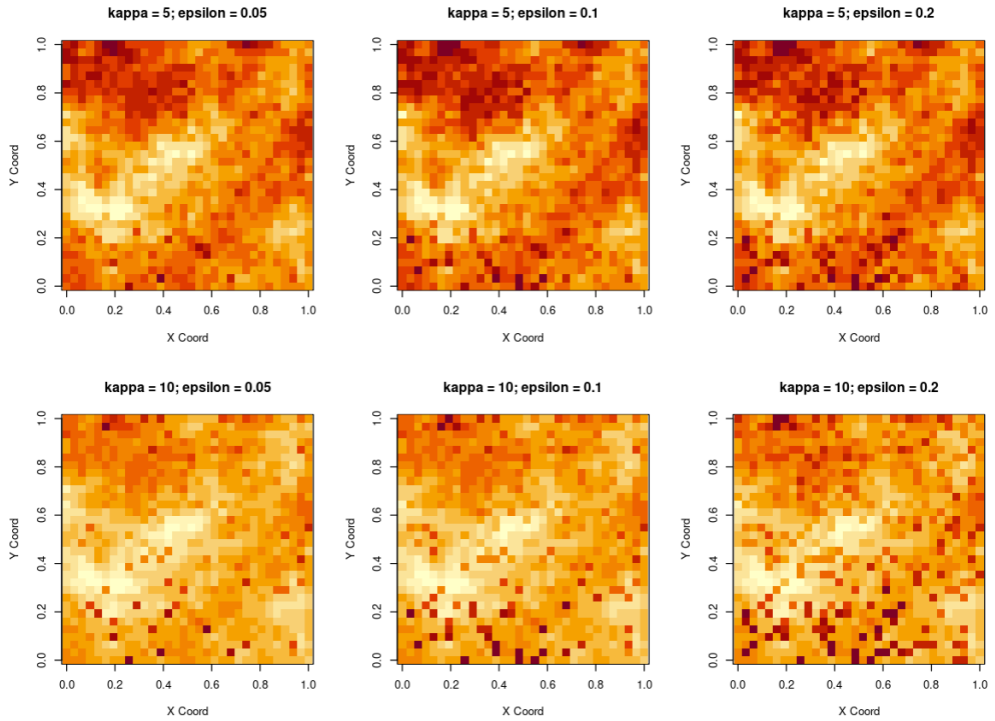
$$E(s) \sim \begin{cases} Gaus(0, c_0), & \textit{with probability } 1 - \epsilon \\ Gaus(0, k^2 c_0), & \textit{with probability } \epsilon \end{cases}$$

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 simulation is based on 1000 spatial points on a regular grid, with a mean of 0 and a covariance function with parameters $\sigma^2 = 1$ and $\phi = 0.25$. The base scenario, that is no contamination, is simulated with $\epsilon = 0$.



Then six different contaminated scenarios are simulated based on the combinations of $epsilon = (0.05, 0.1, 0.2)$ and $kappa = (5, 10)$, to assess the impact of contamination on the variogram estimation under different circumstances.

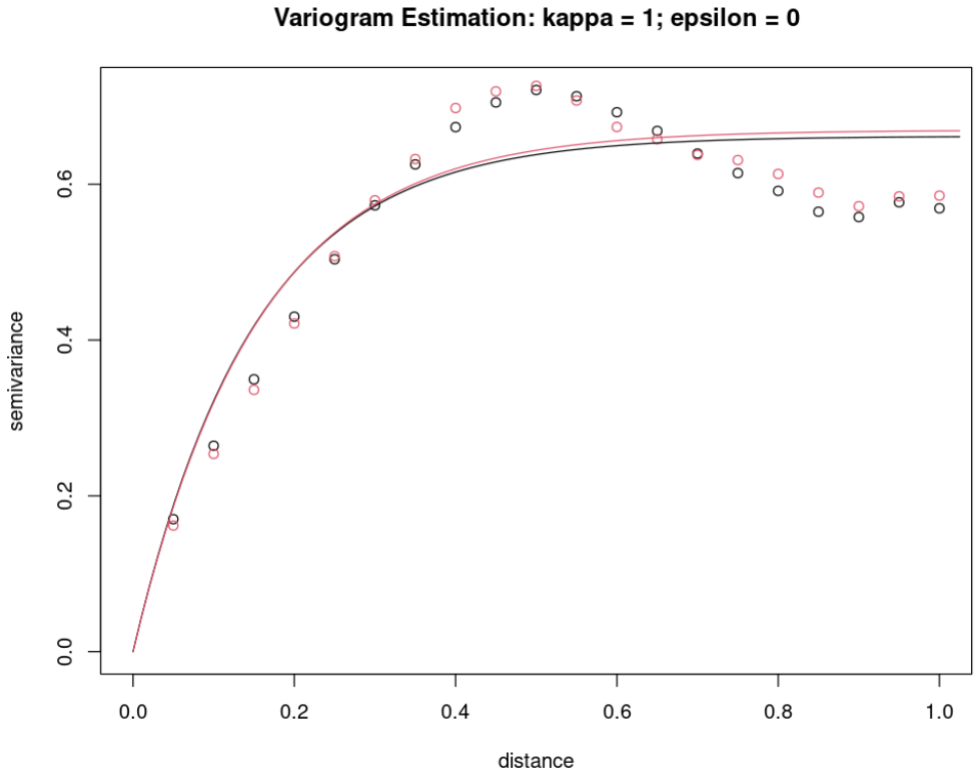


Nugget and different levels of spatial correlation are not considered in this simulation study but they can be easily added to the simulation process.

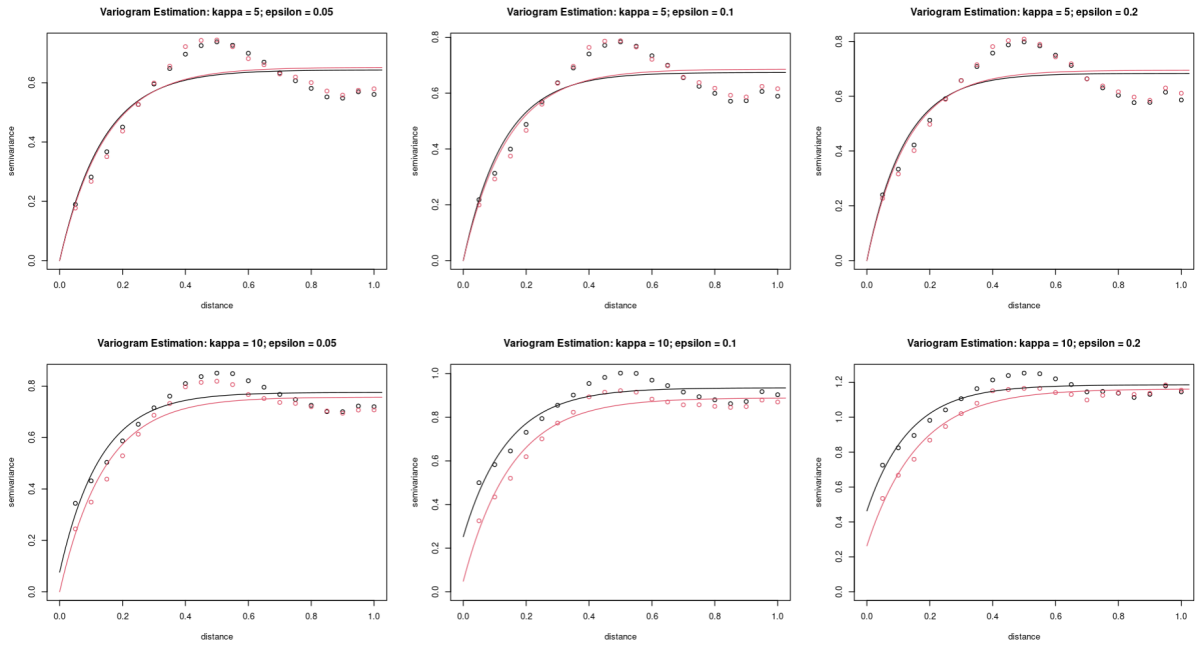
4 Results

The variogram estimation is performed using the classical (in black) and the robust (in red) approaches. The two methods are compared on the different simulated scenarios to assess their robustness to contamination. Moreover, given that outliers are almost always problematic for kriging, a variogram model is also estimated, using the OLS estimator, to see how the outliers affect the model estimates. The estimation process is performed using the functions variog and variofit of the geoR package in R.

The base scenario, that is no contamination, shows that the two methods provide almost identical results.



In the presence of contamination, instead, the two methods provide different results. If the results are analyzed in terms of the outliers' scale (*kappa*), it is possible to observe two different behaviours. For small outliers' scale, that is $kappa = 5$, the two methods provide similar estimations independently of the number of outliers (ϵ). However, for large outliers' scale, that is $kappa = 10$, the estimates are more different the higher the number of outliers.



As expected, from theoretical considerations, the robust variogram estimates are generally smaller than the classical ones. Indeed, the classical approach is heavily influenced primarily by the scale of the outliers, while the robust approach is less sensitive to the contamination. Moreover, the scale of the contamination also strongly affects the estimation of the nugget (which should be always 0), but the robust approach is much less sensitive to this issue.

process	kappa	epsilon	nugget	sigmasq	phi	robust_nugget	robust_sigmasq	robust_phi
GRF	1	0.00	0.00	0.64	0.15	0.00	0.65	0.16
CGRF	5	0.05	0.00	0.64	0.14	0.00	0.65	0.14
CGRF	5	0.10	0.00	0.67	0.13	0.00	0.69	0.14
CGRF	5	0.20	0.00	0.68	0.12	0.00	0.69	0.13
CGRF	10	0.05	0.08	0.70	0.13	0.00	0.76	0.14
CGRF	10	0.10	0.25	0.68	0.14	0.05	0.84	0.15
CGRF	10	0.20	0.46	0.72	0.14	0.26	0.90	0.16

5 Conclusion

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 stable estimates when the scale of the contamination increases. However, if the scale of the contamination is small, then the two methods provide similar results.

References

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.

Journel, A. G. & Huijbregts, C. H. J. (1978), 'Mining geostatistics', *Academic Press, LondonLark RM (2000) Regression analysis with spatially autocorrelated error: simulation studies and application to mapping of soil organic matter. Int J Geogr Inf Sci* **14**(3), 161–195, Section 2.5.

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

Tobler, W. R. (1969), 'Geographical filters and their inverses', *Geographical Analysis* **1**(3), 234–253.