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: spatial statistics, outliers, variogram

1 Introduction

an introductory section describing the main task tackled in the paper (along with relevant bibliographic references)

2 Methods

a methodological section covering methods

3 Simulation

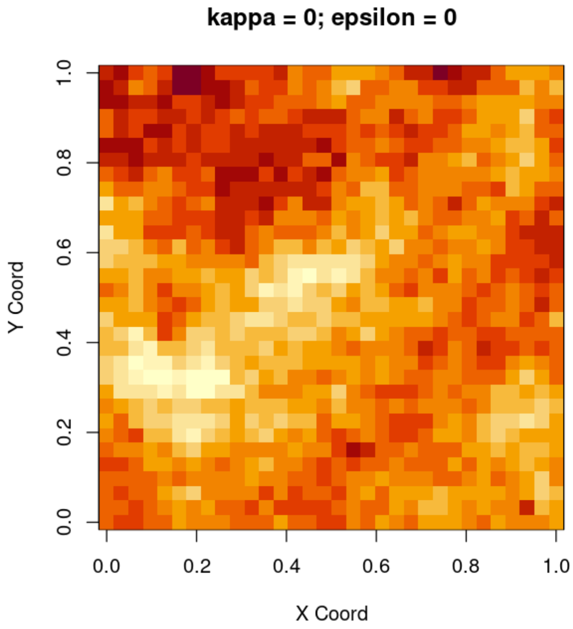
In order to assess the impact of outliers on the variogram estimation based on the classical and the robust approaches, a simulation study is permformed. Simulating Gaussian Random Fields (GRF) and Contaminated Gaussian Random Fields (CGRF) involves different approaches due to the added complexity of 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 Abramowitz (2008), the departure from Gaussianity is obtained simulating a GRF with probability $1 - \epsilon$ and a CGRF with probability ϵ.

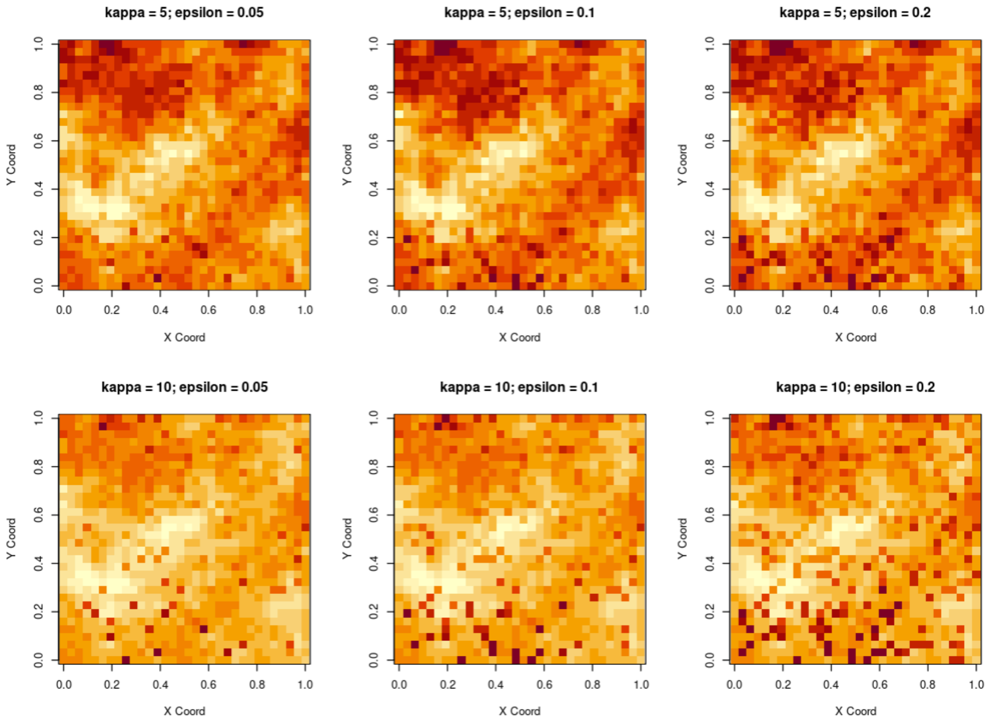
$$\begin{cases} N(0, \sigma^2), & \textit{with probability } 1 - \epsilon \\ N(0, k^2\sigma^2), & \textit{with probability } \epsilon \end{cases}$$

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.

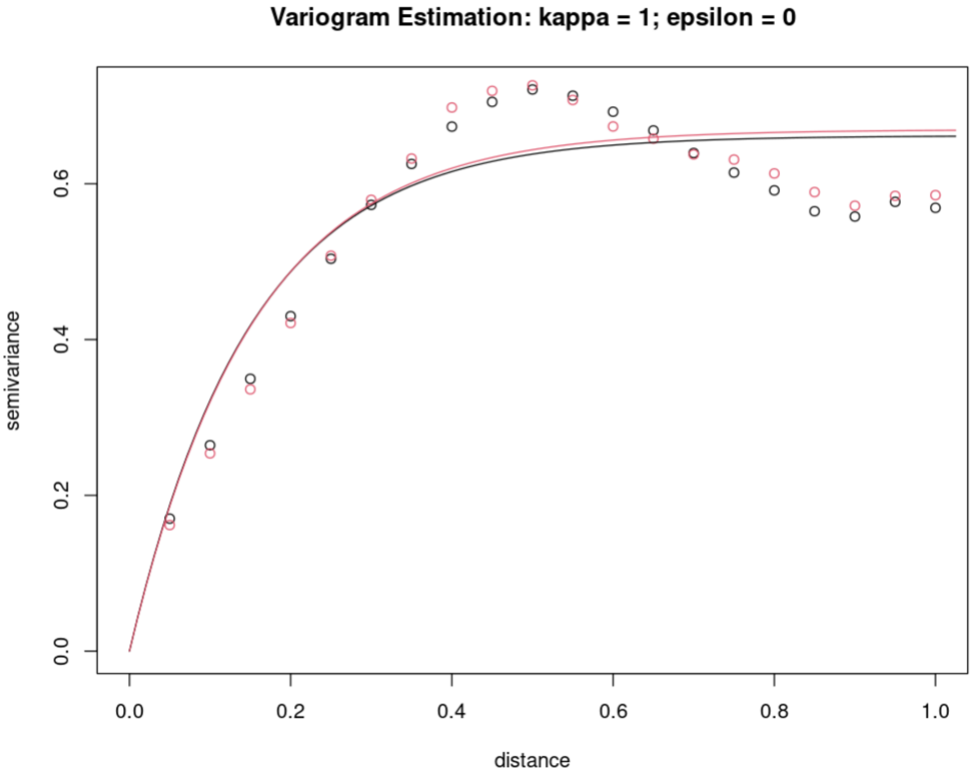


Nugget and different levels of spatial correlation are not considered in this simulation study but they can be easily added to the simulation process.

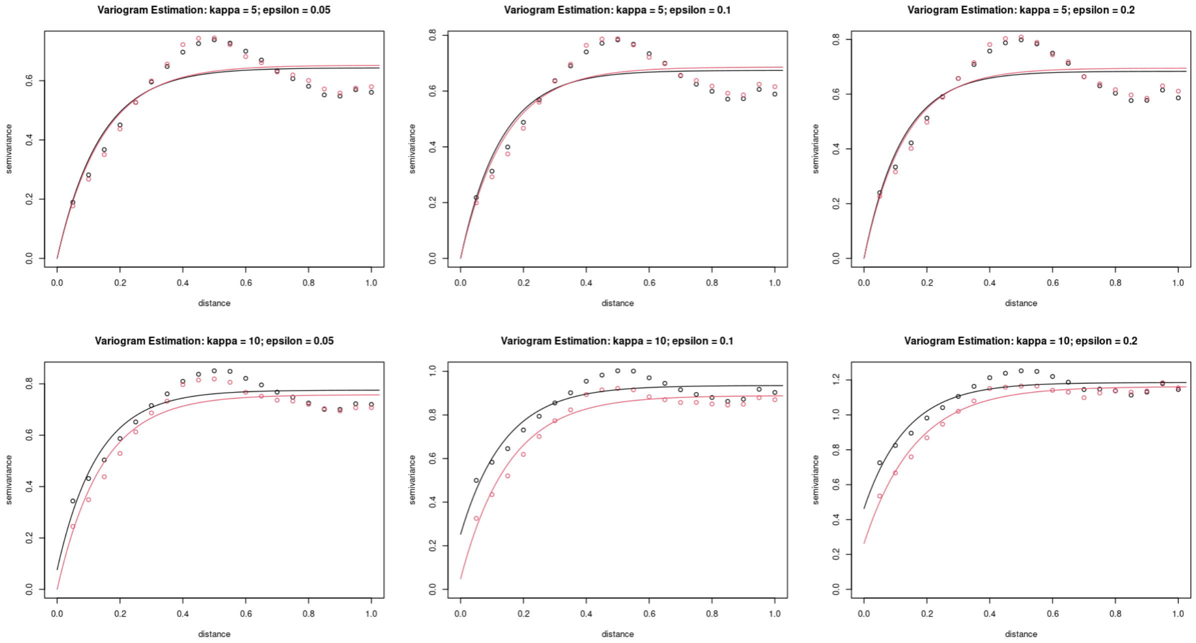
4 Results

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.

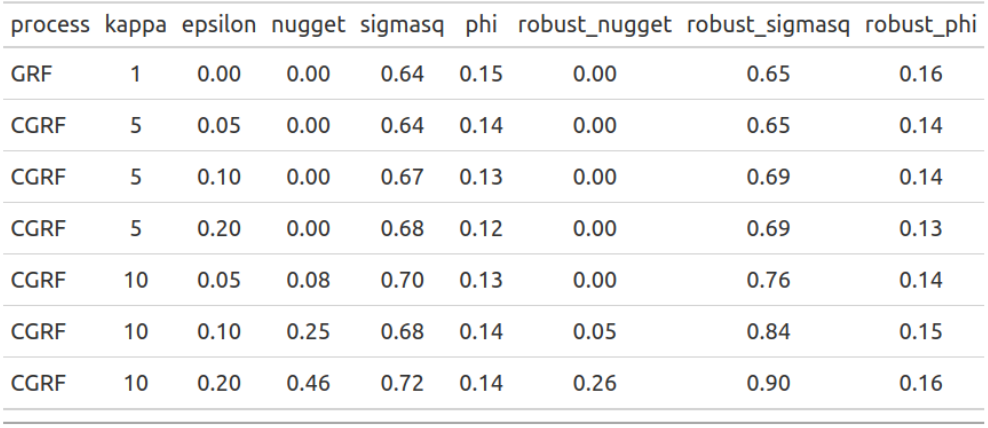
The base scenario, that is no contamination, shows that the two methods provide almost identical results.



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

Abramowitz, A. I. (2008), ‘Forecasting the 2008 Presidential Election with the Time-for-Change Model.’, *PS: Political Science and Politics* .