Trimmed spatio-temporal variogram estimator

Alfonso García-Pérez

Versión final publicada en *Advances in Intelligent Systems and Computing. Bulding Bridges between Soft and Statisical Methodologies for Data Sciencies*, Editores: **L.A. García-Escudero, A. Gordaliza, A. Mayo, M.A. Lubiano Gómez, M.A. Gil, P. Grzegorzewski y O. Hryniewicz, Editorial Springer-Verlag, 2023:174-179.**

https://link.springer.com/chapter/10.1007/978-3-031-15509-3_23

Departamento de Estadística, I.O. y C.N., Universidad Nacional de Educación a Distancia (UNED),

Paseo Senda del Rey 9, 28040 Madrid, Spain; agar-per@ccia.uned.es

Abstract: The spatio-temporal variogram is the key element in spatiotemporal prediction based on kriging, but the classical estimator of this parameter is very sensitive to outliers. In this contributed paper we propose a trimmed estimator of the spatio-temporal variogram as a robust estimator. We obtain an accurate approximation of its distribution with small samples sizes and a scale contaminated normal model.We conclude with an example with real data.

1 Introduction

Let us suppose that we have a spatio-temporal random field $Z(\mathbf{s}, t)$, $(\mathbf{s}, t) \in D \times T$, where $D \subset \mathbb{R}^d$ and $T \subset \mathbb{R}$, which is intrinsically stationary in space and time, i.e., with zero mean in their increments in space and time, and with variance that depends only on displacements in space and differences in time.

The parameter in which we are interested in this paper is the spatio-temporal variogram of Z, defined as

$$2\gamma_Z(\mathbf{h}; \tau) = var(Z(\mathbf{s} + \mathbf{h}; t + \tau) - Z(\mathbf{s}; t))$$

where *var* is the variance of Z, $\mathbf{h}$ a spatial lag and τ a temporal lag. Furthermore, we shall assume that Z is spatially isotropic, i.e., that the variogram depends on the spatial lag $\mathbf{h}$ only through the Euclidean norm $||\mathbf{h}||$.

To estimate $2\gamma_Z(\mathbf{h}; \tau)$, we shall consider observations Z_u, $u = 1, ..., n$, of $Z(\mathbf{s}, t)$ at spatial locations $\{\mathbf{s}_i : i = 1, ..., m\}$ and time moments $\{t_j : j = 1, ..., T\}$, where $n = m \cdot T$ is the sample size. The spatio-temporal variogram is usually estimated with the classical method-of-moments estimator, also called empirical spatio-temporal variogram, (Wikle et al. 2019; Varouchakis and Hristopulos 2019; Cressie 1993), $2\_\gamma_Z(\mathbf{h}; \tau)$, where $N_\mathbf{s}(\mathbf{h})$ refers to the set containing all pairs of spatial locations with spatial lag $\mathbf{h}$ and $N_t(\tau)$ refers to the set containing all pairs of time points with time lag τ. Also, $|N(\cdot)|$ will refer to the number of elements in the set $N(\cdot)$.

Let us observe that this estimator is a sample mean of $n(\mathbf{h},\tau) = |N_{\mathbf{s}}(\mathbf{h})| \cdot |N_t(\tau)|$ terms and therefore, very sensitive to outliers.

In García-Pérez (2020) robust estimators of the spatial variogram and accurate approximations for their distributions were obtained. In García-Pérez (2021) these results were extended to the multivariate case with robust estimators for the cross-variogram. In García-Pérez (2022b) the temporal component was included, obtaining robust M-estimators of the spatio-temporal variogram. In this paper we propose, in Sect. 2, α-trimmed estimators of the spatio-temporal variogram. In Sect. 3 we obtain accurate approximations for the distribution of these new estimators. We conclude the paper, in Sect. 4, with a real-world application.

2 α-Trimmed Spatio-Temporal Variogram Estimator

All over the paper we shall assume that the observations come from a scale contaminated normal model (Huber and Ronchetti 2009, p. 2).

$$(1-\epsilon)N(\mu,\sigma^2) + \epsilon N(\mu, g^2\sigma^2)$$

$\epsilon \in (0,1)$ and $g > 1$. This class of distributions is considered the usual model class in robustness studies because it establishes a neighborhood of the standard model distribution, the *contamination neighborhood*, within which the underlying model lies (Huber and Ronchetti 2009, p. 12).

Let us consider the transformation $X_{ij} = (Z(\mathbf{s}_i + \mathbf{h}; t_j + \tau) - Z(\mathbf{s}_i; t_j))^2$, $\forall \mathbf{s}_i, t_j$. These new variables will be shortened by X_u, $u = 1, ...n$, and will be considered as a sample of a new variable $X = (Z(\mathbf{s} + \mathbf{h}; t + \tau) - Z(\mathbf{s}; t))^2$, defined from the lags of Z in space and time. Now, the parameter of interest is $2\,\gamma_z(\mathbf{h};\tau) = E[X]$, and the problem proposed in the paper is now the problem of estimating the expectation of the random variable X, obtained from the original Z through this transformation.

If can accept a linear semivariogram for the n original observations Z_u and linear cross-variograms for each pair (Z_i, Z_k), then we can admit independence in the X_u (García-Pérez 2022b, Sect. 4).

Considering a scale contaminated normal model for the original Z_u observations, the distribution of the transformed variables X_u is, (García-Pérez 2022b, Sect. 2.3)

$$F = (1-\epsilon)\,2\,\gamma_z(\mathbf{h};\tau)\,\chi^2_1 + \epsilon\, g^2\, 2\,\gamma_z(\mathbf{h};\tau)\,\chi^2_1$$

where χ^2_1 is a Chi-Square distribution with 1 degree of freedom.

When we think in trimming the data we have, mainly, two possibilities: First, to consider all the X_u, $u = 1, ..., n(\mathbf{h},\tau)$, observations as homogenous to be trimmed, or second, to trim by time moments.

From a robustness point of view, trimming by time moments can hide outliers if there is a time moment with many of them; so, it is worse than considering all the observations at once. Hence, we shall choose the first option.

For this reason, we define the *α-trimmed spatio-temporal variogram estimator*, $2\widehat{\gamma}_\alpha(\mathbf{h};\tau)$, defined again for the transformed $n(\mathbf{h},\tau)$ variables X_u as follows:

If we trim the $100\cdot\alpha\%$ of the smallest and the $100\cdot\alpha\%$ of the largest ordered data $X_{(u)}$, the (symmetrically) sample *α-trimmed spatio-temporal variogram estimator* is defined as

$$2\widehat{\gamma}_\alpha(\mathbf{h};\tau)=\frac{1}{n(\mathbf{h},\tau)-2r}\left(X_{(r+1)}+...+X_{(n(\mathbf{h},\tau)-r)}\right)=\overline{X}_\alpha$$

where $r=[n(\mathbf{h},\tau)\alpha]$ if $[\,.\,]$ stands for the integer part.

An asymmetric trimmed spatio-temporal variogram estimator could also be a good option because the observations X_u are positive, since they come from squared differences of the original Z_u.

3 VOM+SAD Approximation of the Distribution of the α-Trimmed Spatio-Temporal Variogram Estimator

Obtaining the distribution of the estimator is necessary to be able to assess its statistical properties and, specifically, its robustness properties. In addition, we can make robust inferences with it such as intervals and robust tests. Moreover, knowing its distribution is useful to reduce the number of temporal lags, as we do in García-Pérez (2022b, Sect. 8) for M-estimators. Even, it would be possible to choose a variogram model in a robust way, as we do in García-Pérez (2022a).

With a von Mises expansion (Von Mises 1947) we can obtain an approximation (VOM approximation) for the distribution of the estimator under a contaminated normal model, computing the approximation under just a normal model, doing in this way the problem easier. This approximation depends on the Hampel's influence function of the tail probability functional. If we have a small sample size, we can use a saddlepoint approximation (SAD approximation) to approximate this functional. Combining both approximations we obtain a VOM+SAD approximation for the distribution of the estimator.

An accurate VOM+SAD approximation of the distribution of the sample α-trimmed mean is obtained in García-Pérez (2016). We see there we can base our approximation for the trimmed mean distribution on an approximation for the classical sample mean. Hence, we can approximate the small sample distribution

of the α-trimmed spatio-temporal variogram estimator, using the small sample distribution of the empirical spatio-temporal estimator, $P_F\{2\widehat{\gamma}_z(\mathbf{h};\tau) > a\}$. Namely, if the number of iterations k is large and also the approximation stabilizes when we increase k, the approximation is

$$P_F\{2\widehat{\gamma}_\alpha(\mathbf{h};\tau) > a\} \simeq (1+n(\mathbf{h},\tau)\,c_1)^{k+1}\,(1+n(\mathbf{h},\tau)\,c_2)^{k+1}\,P_F\{2\widehat{\gamma}_z(\mathbf{h};\tau) > a\}$$

where $c_1 = \left[(1-2\alpha)^{1/(k+1)} - 1\right]$ and $c_2 = \left[1/(1-2\alpha)^{1/(k+1)} - 1\right]$.

In García-Pérez (2022b) an accurate approximation for the distribution of the empirical spatio-temporal estimator is obtained; hence, an accurate approximation for the tail probability of the sample α-trimmed spatio-temporal variogram estimator $2\widehat{\gamma}_\alpha(\mathbf{h};\tau)$, under a scale contaminated normal model, is

$$\begin{aligned} P_F\{2\widehat{\gamma}_\alpha(\mathbf{h};\tau) > a\} \simeq & (1+n(\mathbf{h},\tau)\,c_1)^{k+1}\,(1+n(\mathbf{h},\tau)\,c_2)^{k+1}\left[P\left\{\chi^2_{n(\mathbf{h},\tau)} > \frac{a\,n(\mathbf{h},\tau)}{2\gamma_z(\mathbf{h};\tau)}\right\}\right. \\ & + \epsilon\sqrt{n(\mathbf{h},\tau)}\,\frac{2\gamma_z(\mathbf{h};\tau)}{\sqrt{\pi}(a-2\gamma_z(\mathbf{h};\tau))} \\ & \cdot \exp\left\{-\frac{n(\mathbf{h},\tau)}{2}\left(\frac{a}{2\gamma_z(\mathbf{h};\tau)} - 1 - \log\frac{a}{2\gamma_z(\mathbf{h};\tau)}\right)\right\} \\ & \left.\cdot\left(\frac{\sqrt{2\gamma_z(\mathbf{h};\tau)}}{\sqrt{a-ag^2+2\,g^2\gamma_z(\mathbf{h};\tau)}} - 1\right)\right]. \end{aligned}$$

4 Example

Let us consider daily weather data, obtained in the US National Oceanic and Atmospheric Administration (NOAA) National Climatic Data Center (Wikle et al. 2019). In this data set we shall consider the variable Tmax, the daily maximum temperature in Fahrenheit degrees.

The values of the Classical Spatio-Temporal Variogram Estimator, the 0.05-trimmed spatio-temporal variogram estimator, the 0.1-trimmed spatio-temporal variogram estimator and the 0.2-trimmed spatio-temporal variogram estimator are given in the Supplementary Material, at https://www2.uned.es/pea-metodos-estadisticos-aplicados/trimmed-spa-temp-variogram.htm.

In Fig. 1 here we plot the values of the 0.1-trimmed spatio-temporal variogram estimator obtained in this example.

It is possible to see in the Supplementary Material, p. 15, that there is some differences between the 0.1-trimmed spatio-temporal variogram estimator and the classical one, at some spatial and temporal lags, differences that could be attributed to outliers.

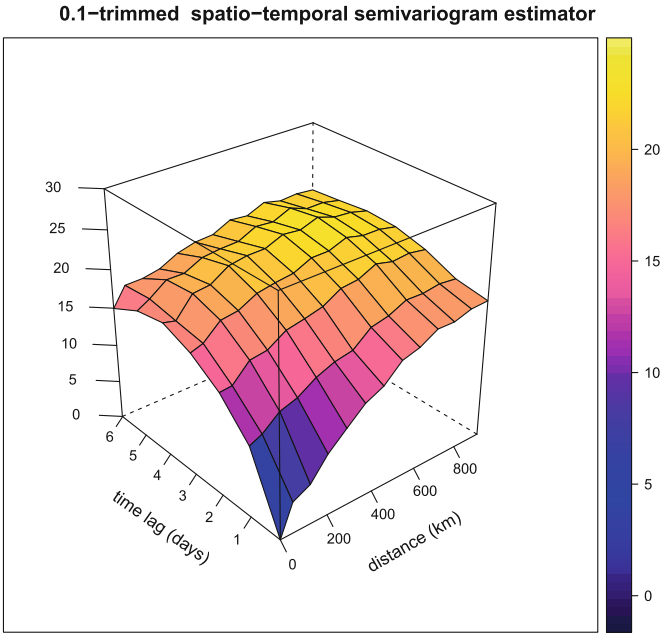

Fig. 1. Three-dimensional picture of the 0.1-trimmed spatio-temporal semivariogram estimator of daily Tmax from the NOAA data during July 2003

5 Conclusions and Further Research

In this paper we define a new robust Trimmed Spatio-Temporal Variogram Estimator and we give an accurate approximation of its distribution. As further research we think that, with this approximation, we could test if it is possible to reduce the number of temporal lags. We also think that it would be possible to obtain an approximation to the distribution of the difference of two α-trimmed spatio-temporal variogram estimators with which we could detect spatio-temporal outliers as it is done in García-Pérez (2022b) for M-estimators.

Acknowledgements. The author is very grateful to the referee and to the *Ministerio de Ciencia e Innovación*.

References

Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)
García-Pérez, A.: A von Mises approximation to the small sample distribution of the trimmed mean. Metrika **79**(4), 369–388 (2016)
García-Pérez, A.: Saddlepoint approximations for the distribution of some robust estimators of the variogram. Metrika **83**, 69–91 (2020)
García-Pérez, A.: New robust cross-variogram estimators and approximations for their distributions based on saddlepoint techniques. Mathematics **9**, 762 (2021)

García-Pérez, A.: Variogram model selection. In: Balakrishnan, N., Gil, M.A., Martin, N., Morales, D., Pardo, M.C. (eds.) Trends in Mathematical, Information and Data Sciences, Studies in Systems, Decision and Control, vol. 445. Springer, Heidelberg (2022a). https://doi.org/10.1007/978-3-031-04137-2_3

García-Pérez, A.: On robustness for spatio-temporal data. Mathematics **10**, 1785 (2022)

Huber, P.J., Ronchetti, E.M.: Robust Statistics, 2nd edn. Wiley, New York (2009)

Varouchakis, E.A., Hristopulos, D.T.: Comparison of spatiotemporal variogram functions based on a sparse dataset of groundwater level variations. Spat. Stat. **34**, 1–18 (2019)

von Mises, R.: On the asymptotic distribution of differentiable statistical functions. Ann. Math. Stat. **18**, 309–348 (1947)

Wikle, C.K., Zammit-Mangion, A., Cressie, N.: Spatio-Temporal Statistics with R. Chapman & Hall/CRC, New York (2019)