Variogram model selection

Alfonso García-Pérez

Versión final publicada en *Trends in Mathematical, Information and Data Sciences*, Editores: **N. Balakrishnan, M.A. Gil, N. Martín, D. Morales y M.C. Pardo, Editorial Springer-Verlag, 2023:21-27.**

https://link.springer.com/chapter/10.1007/978-3-031-04137-2_3

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: A common problem in geostatistics is variogram estimation, in order to choose an acceptable model for kriging. Nevertheless, there is no standard method, first, to test if a particular model can be accepted as valid and, second, to choose among several competing variogram models. The problem is even more complex if, in addition, there are outliers in the data. In this paper we propose to use the distribution of some classical and robust variogram estimators to test, first, the validity of a particular model, accepting it if the *p*-value of the test, with this particular model as null hypothesis, is large enough and, second, to compare several competing models, choosing the model with the largest *p*-value among several acceptable models.

Keywords: Robustness; Spatial data; saddlepoint approximations

MSC: 62F35; 62E17; 62H11

1 Introduction

A common problem in geostatistics is variogram model selection among several competing models, after the variogram has been estimated, usually by weighted least squares.
Among all the models that apparently fit well, you might choose from among them the one with smallest residual sum of squares, or smallest mean square, or the usual Matheron's estimator [8].
Sometimes, the chosen model is the one with smallest Akaike's information criterion (AIC) [1]

$$\text{AIC} = -2\log(\text{maximized likelihood}) + 2p$$

being p the number of parameters of the model.

AIC is usually estimated by

$$\widehat{\text{AIC}} = \{n \log(2\pi/n) + n + 2\} + n \log R + 2p$$

where n is the number of points on the variogram, and R is the mean of the squared residuals between observed values and the fitted model (Webster and Oliver [13], p. 105).

Because the first term is constant, the model with the smallest $n \log R + 2p$ is chosen.

Combining both criteria, the model with the smallest mean squared residual and the smallest $n \log R + 2p$ is, usually, the selected model.

But the chosen model might not be significant enough because there is no probability distribution to compare with.

In this sense, Webster and McBratney [11] propose an F test for nested models, and suggest other possible criteria.

In this context, equations for estimating the estimation variances for variograms (with a bounded sill) are given in Matheron [9] and in Muñoz-Pardo [10], solving them by numerical integration. Also, Webster and Oliver [12] obtain confidence limits by Monte Carlo methods.

These results are valid considering only classical estimators and observations with a normal model distribution. The paper by Gorsich and Genton [6] has these purposes from a nonparametric point of view.

In García-Pérez [3] approximated distributions, under a contaminated normal model, for classical and robust variogram estimators are obtained. The aim of this paper is to use these approximations for the distributions of the Matheron's estimator and some robust ones, first, to valid a particular variogram model and, then, to compare among several competing variogram models.

2 Robust Estimators of the Variogram

Let us suppose that an univariate random variable Z is observed at some known fixed locations $\mathbf{s}_i \in D$, being D a fixed subset of $\mathbb{R}^d$, $d \geq 1$, and let us assume that the variable Z satisfies the intrinsic stationarity property, i.e., the differences have zero mean

$$E[Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})] = 0, \quad \forall \mathbf{s}, \mathbf{s}+\mathbf{h} \in D$$

and the variance depends only on lag $\mathbf{h}$,

$$V(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = 2\,\gamma(\mathbf{h}), \quad \forall \mathbf{s}, \mathbf{s}+\mathbf{h} \in D,$$

being the function

$$2\,\gamma(\mathbf{h}) = V(Z(\mathbf{s}+\mathbf{h}) - Z(\mathbf{s})) = E[(Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h}))^2]$$

the variogram. This is estimated with the classical Matheron's estimator,

$$2\hat{\gamma}_M(\mathbf{h}) = \frac{1}{N_h}\sum_{i=1}^{N_h}(Z_{i+h} - Z_i)^2$$

where the sample size $n = N_h$ is the cardinality of $N(\mathbf{h}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{h}\}$. In García-Pérez [3] some robust estimators of the variogram were introduced.

If we transform the original observations Z_i by $Y_i = (Z_{i+h} - Z_i)^2$, robust M-estimators T_n of the variogram can be obtained as solutions of the equation

$$\sum_{i=1}^{n}\psi(Y_i, T_n) = 0.$$

If a linearized variogram can be accepted, the transformed variables Y_i can be considered as independent.

If we assume a scale contaminated normal model,

$$F = (1-\varepsilon)\,N(\mu, \sigma) + \varepsilon\,N(\mu, g\sigma)$$

with $\varepsilon \in (0, 1)$ (usually small) and $g > 1$, for the marginal distributions of the original observations Z_i, that means a distribution $F = (1-\varepsilon)\,2\,\gamma(\mathbf{h})\,\chi_1^2 + \varepsilon\,g^2\,2\,\gamma(\mathbf{h})\,\chi_1^2$ for the transformed observations Y_i, in García-Pérez [3] it is proved that a saddlepoint approximation (VOM+SAD) for the distribution of T_n is

$$P_F\{T_n > t\} \simeq P_G\{T_n > t\} + \varepsilon\,\frac{\phi(s)}{r_1}\sqrt{n}\left(\frac{\int e^{z_0\psi(x,t)}dH(x)}{\int e^{z_0\psi(y,t)}dG(y)} - 1\right) \qquad (1)$$

being $G = 2\gamma(\mathbf{h})\chi_1^2$, $H = g^2 2\gamma(\mathbf{h})\chi_1^2$, ϕ the density function of the standard normal distribution, s and r_1 are the functionals

$$s = \sqrt{-2nK(z_0, t)}, \qquad r_1 = z_0\sqrt{K''(z_0, t)}$$

$K(\lambda, t)$ the function

$$K(\lambda, t) = \log\int_{-\infty}^{\infty} e^{\lambda\psi(y,t)}\,dG(y)$$

$K''(\lambda, t)$ ($K'(\lambda, t)$) the second (the first) partial derivative of $K(\lambda, t)$ with respect to the first variable and z_0 the saddlepoint, i.e., the solution of the *saddlepoint equation*

$$K'(z_0, t) = \int_{-\infty}^{\infty} e^{z_0\psi(y,t)}\,\psi(y, t)\,dG(y) = 0.$$

Approximation (1) is easy to compute with R for the Matheron's estimator and for robust M-estimators, as it is explained in García-Pérez [3]. In that paper, an α-trimmed variogram estimator is also introduced and its VOM+SAD distribution, obtained.

3 Acceptance of a Model and Variogram Model Comparison

The VOM+SAD approximations obtained in García-Pérez [3] for classical and robust variogram estimators, can be used to test if a particular variogram model $2\gamma(\mathbf{h})$ can be accepted to explain a variogram estimator $T_n = 2\hat{\gamma}(\mathbf{h})$ and also, to compare several variogram models.

Let us assume model $2\gamma(\mathbf{h})$ as null hypothesis and $2\hat{\gamma}(\mathbf{h})$ as a variogram estimator. We consider the test statistic

$$S_n = \sup_{\mathbf{h}} \left\| 2\hat{\gamma}(\mathbf{h}) - 2\gamma(\mathbf{h}) \right\| = \max_{1 \le ||\mathbf{h}|| \le k} \left\| 2\hat{\gamma}(\mathbf{h}) - 2\gamma(\mathbf{h}) \right\|$$

taking values s_n, assuming there are k lags.

If the p-value of this test,

$$P\{S_n > s_n\}$$

is large enough, the model will be accepted; otherwise the model will be rejected.

If several competing models are accepted, the model for which this p-value is the largest, will be the selected one.

In García-Pérez [3], it is obtained that the cumulative distribution function of S_n

$$F_{S_n}(s_n) = 1 - P\{S_n > s_n\}$$

is

$$F_{S_n}(s_n) = \prod_{||\mathbf{h}||=1}^{k} \left[P_{2\gamma(h)}\{2\hat{\gamma}(\mathbf{h}) > -s_n + 2\gamma(\mathbf{h})\} - P_{2\gamma(h)}\{2\hat{\gamma}(\mathbf{h}) > s_n + 2\gamma(\mathbf{h})\} \right]$$

being this tail probabilities computed with the VOM+SAD approximations.

4 Example

Let us consider log Calcium data (mg/l), one of the eight variables observed in the groundwater data analysis around the city of Madrakah, a town located in the Wadi Usfan region in western Saudi Arabia, (Marko et al. [7]). In Cabrero-Ortega and

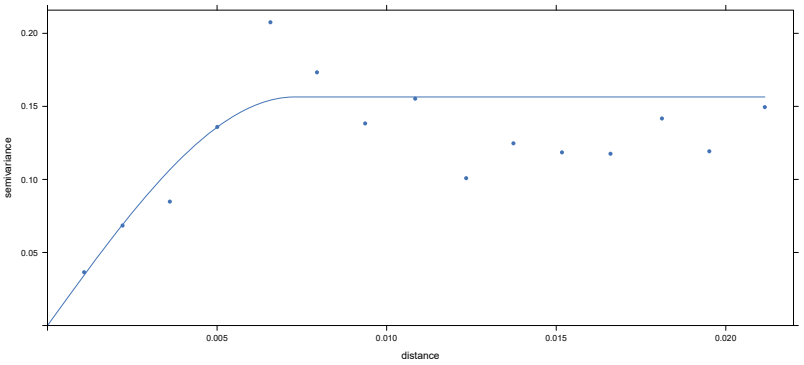

Fig. 1 Matheron's estimator and a Spherical model

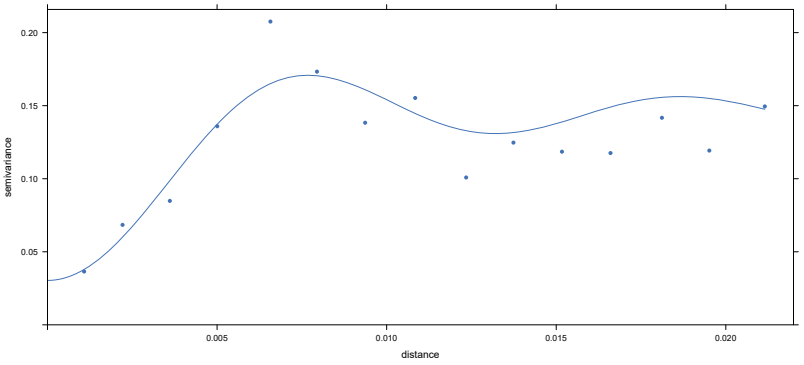

Fig. 2 Matheron's estimator and a Cardinal Sine model

García-Pérez ([2], pp. 303–310), a classical methodology is applied to these data, concluding that an Spherical model with partial sill = 0.1564478, nugget = 0 and range = 0.007289068 is suitable, see Fig. 1. We also observe Matheron's estimations for several lags in this figure and some outliers, appreciating that these estimates seem to be affected by them.

In García-Pérez [3], we define robust estimates for these data and we also prove that the linearized versions of the variogram models (classical, 0.05-trimmed and Huber) can be accepted. Hence, we can consider the transformed observations Y_i as independent. We also obtain the VOM+SAD approximations for their distributions.

But let us observe that, in Cabrero-Ortega and García-Pérez [2], we also mention that a Cardinal Sine model with partial sill = 0.11533833, nugget = 0.03038008 and range = 0.005372508, can also be accepted for these data, as we see in Fig. 2.

We check now if both models have p-values large enough to be accepted and which one is the largest. The p-values for the Spherical model, computed with the VOM+SAD approximations are included in the middle-hand of Table 1. In the right-

Table 1 P-values for the Spherical model (middle) and Cardinal Sine model (right), considering the classical Matheron's estimator and two robust ones

Estimator\Model	Spherical model	Cardinal Sine model
Classical	0.2270516	0.0001244
0.05-trimmed	0.1333519	0.0862922
Huber	0.0157108	0.1036955

hand of the table we show the p-values for the Cardinal Sine model. The computations are in the Supplementary Material available on the website

https://www2.uned.es/pea-metodos-estadisticos-aplicados/VariogramSelection.htm

Although both models are accepted using the standard criteria, from this table we see that Cardinal Sine model cannot be accepted considering the distribution of Matheron's estimator and that Spherical model can be accepted.

Nevertheless, if we use robust methods, the conclusion is the opposite one because of the outliers in the data: first, with the 0.05-trimmed variogram estimator both models are acceptable but, because of the asymmetry, it is better to use Huber's estimator with which we conclude that Cardinal Sine model should be the selected one.

5 Conclusions and Future Works

The selection of a valid variogram model is a key question in geostatistics. In this paper we propose to establish a test to do this, in which the null hypothesis is the suggested variogram model, which is accepted if the p-value is large enough.

If several model are valid, we propose to chose the model with the largest p-value.

This proposal is especially useful when there are outliers in the data set because robust variogram estimators can be used in the proposal.

This test is performed with the VOM+SAD approximations to the distribution of the classical a robust variogram estimators obtained in García-Pérez [3].

These ideas can be extended to the multivariate situation through the cross-variogram, following the results obtained in García-Pérez [4] and, even, to the spatio-temporal framework, with the results developed in García-Pérez [5].

Acknowledgements This work is partially supported by Grant PGC2018-095194-B-I00 from Ministerio de Ciencia, Innovación y Universidades (Spain).

References

1. Akaike, H.: Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csáki, F. (eds.) 2nd International Symposium on Information Theory, pp. 267–281. Akadémiai Kiadó, Budapest (1973)
2. Cabrero-Ortega, M.Y., García-Pérez, A.: Análisis estadístico de datos espaciales con QGIS y R. Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain (2020)
3. García-Pérez, A.: Saddlepoint approximations for the distribution of some robust estimators of the variogram. Metrika **83**(1), 69–91 (2020)
4. García-Pérez, A.: New robust cross-variogram estimators and approximations of their distributions based on saddlepoint techniques (2020). Submitted
5. García-Pérez, A.: Robust spatio-temporal variogram estimators and a saddlepoint approximation for their distributions (2020). Submitted
6. Gorsich, D.J., Genton, M.G.: Variogram model selection via nonparametric derivative estimation. Math. Geol. **32**(3), 249–270 (2000)
7. Marko, K., Al-Amri, N.S., Elfeki, A.M.M.: Geostatistical analysis using GIS for mapping groundwater quality: case study in the recharge area of Wadi Usfan, western Saudi Arabia. Arab. J. Geosci. **7**, 5239–5252 (2014)
8. Matheron, G.: Traité de géostatistique appliquée, Tome I: Mémoires du Bureau de Recherches Géologiques et Minières, no. 14, Editions Technip, Paris (1962)
9. Matheron, G.: Les variables régionalisées et leur estimation. Masson, Paris (1965)
10. Muñoz-Pardo, J.F.: Approche géostatistique de la variabilité spatiale des milieux géophysique. MA Thesis, Université de Grenoble et lÍnstitut National Polytechnique de Grenoble (1987)
11. Webster, R., McBratney, A.B.: On the Akaike Information Criterion for choosing models for variograms of soil properties. Eur. J. Soil. Sci. **40**, 493–496 (1989)
12. Webster, R., Oliver, M.A.: Sample adequately to estimate variograms of soil properties. Eur. J. Soil. Sci. **43**, 177–192 (1992)
13. Webster, R., Oliver, M.A.: Geostatistics for Environmental Scientists, 2nd edn. Wiley, Chichester (2007)