NONPARAMETRIC BIAS CORRECTED ESTIMATOR OF THE SEMIVARIOGRAM WITH NON-CONSTANT TREND

Rubén Fernández-Casal <*rfcasal@udc.es*>. **University of A Coruña (SPAIN)**
Mario Francisco-Fernández <*mariofr@udc.es*>. **University of A Coruña (SPAIN)**

Abstract

In spatial statistics, the approximation of the spatial dependence structure of a process, through the estimation of the variogram or the covariogram of the variable under consideration, is an important issue. In this work, under a general spatial model, including a mean or trend function, and without assuming any parametric model for this function and for the dependence structure of the process, a general nonparametric estimator of the variogram function is proposed. The new approach consists in applying an iterative algorithm, using the residuals obtained from a nonparametric local linear estimation of the trend function, jointly with a correction of the bias due to the use of these residuals. A simulation study checks the validity of the presented approaches in practice.

Introduction

- Classical statistical methods only include the large-scale variation and assume independent errors. Geostatistical methods, apart from considering the spatial trend (large-scale variation), also take the spatial correlation (small-scale variation) into account.
- In traditional geostatistical approaches, when the mean is not stationary, the usual dependence estimation method consists in removing the trend and estimate the variogram (or the covariogram) from the residuals.
- The direct use of the residuals, even proceeding in the most efficient way, introduces a bias (e.g. Cressie, 1993, section 3.4.3).
- If the final goal is prediction (kriging), the effect of this bias may be small. However, if the analysis requires an accurate estimation of the small-scale variability of the process (such as hypothesis testing or simulation), the results may be strongly influenced by it.
- Additionally, if a nonparametric fit is used to remove the trend, the magnitude of the bias is usually larger (see e.g. Figure 1).
- In this work, under a complete nonparametric model, a new estimator of the semivariogram with a bias correction is proposed.

Dependence estimation with non-constant mean function

- Assume that $\left\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\right\}$ is a spatial process with:

$$Y(\mathbf{x}) = m(\mathbf{x}) + \varepsilon(\mathbf{x}), \tag{1}$$

where $m(\cdot)$ is the trend function and the error term, ε, is a second order stationary process with zero mean and covariogram $C_\varepsilon(\mathbf{x}_i - \mathbf{x}_j)$.

- Under the universal kriging model: $m(\mathbf{x}) = \sum_{j=0}^{p} f_j(\mathbf{x})\beta_j$, where $\left\{f_j(\cdot) : j = 0, ..., p\right\}$ are known functions (usually polynomials) and $\beta = (\beta_0, ..., \beta_p)^t$ is a unknown vector of coeficients.
- The most efficient estimator in this case is the generalized least squares (gls) estimator, $\hat{\beta}_{gls} = \left(\mathbf{X}^t\Sigma^{-1}\mathbf{X}\right)^{-1}\mathbf{X}^t\Sigma^{-1}\mathbf{Y}$, with $\mathbf{X}$ the design matrix, $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_n))^t$ and Σ the variance-covariance matrix of the errors.
- The covariogram of the error process is usually unknown in practice and, therefore, it has to be estimated. This estimation is generally based on the residuals (see, for instance, Neuman and Jacobson, 1984).
- Even in the most efficient situation, employing the gls residuals $\hat{\varepsilon} = \mathbf{Y} - \mathbf{X}\hat{\beta}_{gls}$, it can be observed that $Var(\hat{\varepsilon}) = \Sigma - \mathbf{X}(\mathbf{X}^t\Sigma^{-1}\mathbf{X})^{-1}\mathbf{X}^t$.
- In the linear case and using a parametric model jointly with the empirical variogram estimator, a correction for this bias was proposed in Beckers and Bogaert (1998).

Nonparametric trend estimation

- In the spatial framework, the classical nonparametric local linear estimator for $m(\cdot)$ at a location $\mathbf{x}$ is the solution for α to the least squares minimization problem

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i=1}^{n} \left\{Y(\mathbf{x}_i) - \alpha - \boldsymbol{\beta}^t(\mathbf{x}_i - \mathbf{x})\right\}^2 K_{\mathbf{H}}(\mathbf{x}_i - \mathbf{x}),$$

where $\mathbf{H}$ is a 2×2 symmetric positive definite matrix; K is a bivariate kernel and $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$.

- The local linear regression estimator can be written explicitly as:

$$\hat{m}_{\mathbf{H}}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{Y} \equiv s_\mathbf{x}^t \mathbf{Y},$$

where $\mathbf{e}_1$ is a vector with 1 in the first entry and all other entries 0, $\mathbf{X}_\mathbf{x}$ is a matrix with ith row equal to $(1, (\mathbf{x}_i - \mathbf{x})^t)$, and $\mathbf{W}_\mathbf{x} = \mathtt{diag}\left\{K_{\mathbf{H}}(\mathbf{x}_1 - \mathbf{x}), ..., K_{\mathbf{H}}(\mathbf{x}_n - \mathbf{x})\right\}$.

- The approach of Francisco-Fernández and Opsomer (2005), that takes the dependence of the data into account, can be used to select the bandwidth $\mathbf{H}$.
- In this case, the nonparametric residuals are given by $\tilde{\varepsilon} = \mathbf{Y} - \mathbf{S}\mathbf{Y}$, with $\mathbf{S}$ the $n \times n$ matrix whose ith row is equal to $\mathbf{s}_{\mathbf{x}_i}^t$, the smoother vector for $\mathbf{x} = \mathbf{x}_i$.
- The estimator of Σ based on those residuals actually turns out to be an estimator of $Var(\tilde{\varepsilon}) = (\mathbf{I} - \mathbf{S})\Sigma(\mathbf{I} - \mathbf{S})^t = \Sigma + \mathbf{B}$, where $\mathbf{B} = \mathbf{S}\Sigma\mathbf{S}^t - \Sigma\mathbf{S}^t - \mathbf{S}\Sigma$ represents the bias.
- A similar expression for the bias is obtained for the variogram:

$$Var\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right) = Var\left(\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j)\right) + b_{ii} + b_{jj} - 2b_{ij}.$$

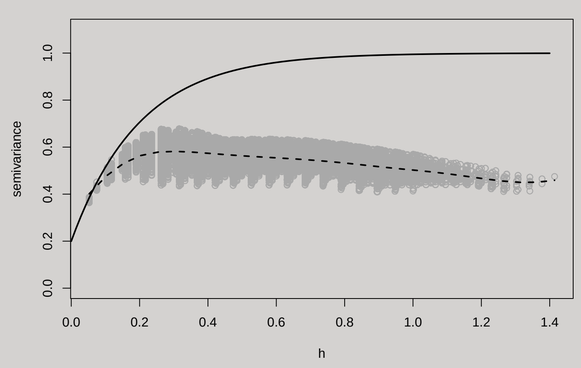

Figure 1.Theoretical semivariogram (solid line) and semivariances of the residuals.

- Additionally, a nonparametric kernel estimator of the semivariogram could be also used. The pilot local linear estimate $\hat{\gamma}(\mathbf{u})$ is the solution for α to the least squares minimization problem:

$$\min_{\alpha,\boldsymbol{\beta}} \sum_{i<j} \left\{\frac{1}{2}\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right)^2 - \alpha - \boldsymbol{\beta}^t\left(\mathbf{x}_i - \mathbf{x}_j - \mathbf{u}\right)\right\}^2 \cdot K_{\mathbf{G}}\left(\mathbf{x}_i - \mathbf{x}_j - \mathbf{u}\right), \tag{2}$$

where $\mathbf{G}$ is the corresponding bandwidth matrix.

Bias correction in the nonparametric case

To obtain a more reliable estimator of the dependence structure, the following method is proposed:

1. Keeping $\mathbf{S}$ fixed, compute the residuals $\tilde{\varepsilon}$. From these residuals, obtain a valid variogram (or covariogram) model, for instance, assuming a parametric model or using the nonparametric approach given in Fernández-Casal et al. (2003). Then, construct prior estimators of Σ and $\mathbf{B}$ (denoted respectively by $\tilde{\Sigma}^{(0)}$ and $\hat{\mathbf{B}}^{(0)}$).
2. At each stage k, using $\hat{\mathbf{B}}^{(k-1)}$, a pilot variogram estimator is obtained, replacing in (2) the differences $\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right)^2$ by

$$\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right)^2 - \hat{b}_{ii}^{(k-1)} - \hat{b}_{jj}^{(k-1)} + 2\hat{b}_{ij}^{(k-1)}.$$

Next, fit a valid variogram model and obtain the corresponding $\tilde{\Sigma}^{(k)}$ and $\hat{\mathbf{B}}^{(k)}$.

3. Repeat step 2 until convergence.

Simulation results

- $N = 1,000$ samples of different sizes were generated following model (1) on a regular grid in the unit square, with mean function $m(x_1, x_2) = \sin(2\pi x_1) + 4(x_2 - 0.5)^2$ and random errors ε_i normally distributed with zero mean and isotropic exponential covariogram:

$$\gamma_\theta(\mathbf{u}) = c_0 + c_1\left(1 - \exp\left(-3\frac{\|\mathbf{u}\|}{r}\right)\right),$$

(for $\mathbf{u} \neq \mathbf{0}$), where c_0 is the nugget effect, c_1 is the partial sill and r is the practical range. The sill (variance) was fixed to $\sigma^2 = 1$ ($c_1 = \sigma^2 - c_0$).

- The values considered in the simulations were:
  - Sample sizes of $n = 10 \times 10$, 15×15 and 20×20.
  - Practical ranges of $r = 0.3$, 0.6 and 0.9.
  - Nugget values of $c_0 = 0$, 0.2 and 0.5 (100%, 80% and 50% of spatial variability, respectively).
- The bandwidth $\mathbf{H}$ was selected minimizing the mean average squared error:

$$MASE(\mathbf{H}) = \frac{1}{n}(\mathbf{S}\mathbf{m} - \mathbf{m})^t(\mathbf{S}\mathbf{m} - \mathbf{m}) + \frac{1}{n}tr\left(\mathbf{S}\Sigma\mathbf{S}^t\right),$$

where $\mathbf{m} = (m(\mathbf{x}_1), ..., m(\mathbf{x}_n))^t$, to avoid the effect of the trend bandwidth selection criterion on the results.

- Following Fernández-Casal et al. (2003), the standard cross-validation method was used to select the bandwidth $\mathbf{G}$.
- Similar results were obtained with the different simulations settings. For the sake of brevity, only results with $n = 400$, $r = 0.6$ and $c_0 = 0.2$ are shown here.
- Figure 2 shows the averaged values and the .25 and .75 pointwise quantile curves of the nonparametric semivariogram estimates (the corresponding theoretical biases are shown in Figure 1). As expected, substantial bias at large lags are observed when the estimation is based on the uncorrected residuals.

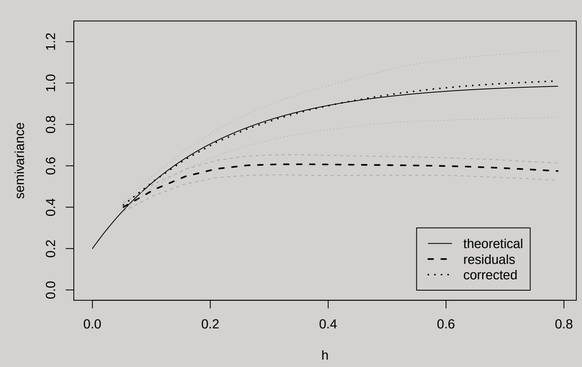

Figure 2. Summaries of the nonparametric semivariogram estimates.

- The bias-corrected estimates are very close to the theoretical semivariogram.
- A summary of the nonparametric semivariogram estimation errors is shown in Table 1, where it is observed the better perfomance of the bias-corrected estimator.

Table 1. Summary of squared errors (SE) and relative squared errors (RSE)

	SE			RSE		
	mean	median	sd	mean	median	sd
Residuals	.086	.081	.069	.100	.100	.071
Corrected	.031	.007	.060	.037	.012	.065

- Additionally, exponential semivariogram models were fitted to the nonparametric pilot estimates by weighted least squares (WLS). Results are shown in Table 2.

Table 2. Summary of parameter estimates obtained by WLS

	$c_0 = 0.2$		$c_1 = 0.8$		$r = 0.6$	
	resid.	corr.	resid.	corr.	resid.	corr.
mean	.215	.247	.394	.861	.629	.940
median	.170	.234	.418	.820	.228	.667
sd	.177	.115	.202	.364	.986	.738

Conclusions

- The direct use of the residuals in the variogram (or the covariogram) estimation may produce a strong underestimation of the small-scale variability of the process. This bias may have a significant impact on the construction of confidence (or prediction) intervals, hypothesis testing or simulation.
- The proposed methodology yields more accurate estimates of the variogram.
- A nonparametric variogram estimate is obtained at the end of the procedure. The proposed algorithm requires the fitting of a valid model. Preliminary tests suggest that these fits have a small impact on the final estimates (simple models may be used). The selection of the final model is left to the user.

Acknowledgments

The research of Mario Francisco-Fernández has been partially supported by Grants MTM2008-00166 (ERDF included) and MTM2011-22392. The research of Rubén Fernández-Casal has been partially supported by MEC Grant MTM2008-03010.

References

1. Beckers, F. and Bogaert, P. (1998). Nonstationary of the mean and unbiased variogram estimation: extension of the weighted least-squares method. *Mathematical Geology* **30**, 223–240.
2. Cressie, N. (1993). *Statistics for Spatial Data*. New York. Wiley.
3. Fernández-Casal, R., González-Manteiga, W. and Febrero-Bande. M. (2003). Space-time dependency modeling using general classes of flexible stationary variogram models. *Journal of Geophysical Research* **108**, 8779.
4. Francisco-Fernández, M. and Opsomer, J. D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *Canadian Journal of Statistics* **33**, 539–558.
5. Neuman, S.P. and Jacobson, E.A. (1984). Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. *Mathematical Geology* **16**, 499–521.