AUTOMATIC NONPARAMETRIC GEOSTATISTICAL MODELING

Rubén Fernández-Casal
University of A Coruña (Spain)
ruben.fcasal@udc.es

Abstract

The modeling of a geostatistical process typically consists in the estimation of the trend and variogram functions. In this work, under a general spatial model and without assuming any parametric form for these functions, a general nonparametric procedure for the modeling of geostatistical data is proposed. The approach consists in an iterative algorithm, combining a local linear estimator of the trend, selecting the bandwidth by a method that takes into account the spatial dependence, and the fit of a Shapiro-Botha variogram model to a set of bias-corrected nonparametric pilot estimates. This algorithm is implemented in the `np.fitgeo` function of the `R` package `npsp` (avaliable on CRAN). .

Introduction

- Assuming that $\left\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\right\}$ is a spatial process that can be modeled as:

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

where $\mu(\cdot)$ is the trend function and the error term ε, is a second order stationary process with zero mean and covariogram $C(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$, with $\mathbf{u} \in D$.

- In this framework, given n observed values $\mathbf{Y} = (Y(\mathbf{x}_1), ..., Y(\mathbf{x}_n))^t$, the goal is to estimate the trend $\mu(\mathbf{x})$ and the semivariogram $\gamma(\mathbf{u}) = C(\mathbf{0}) - C(\mathbf{u})$.

- The geostatistical techniques commonly used to approximate these functions usually assume parametric models, therefore, they can present misspecification problems.

- In this work, under the general spatial model (1), and without assuming any parametric model for the trend function and for the dependence structure of the process, an automatic nonparametric procedure for the modeling of geostatistical data is proposed.

Nonparametric estimation

Local linear trend estimation

- The local linear trend estimator $\hat{\mu}_{\mathbf{H}}(\mathbf{x})$ (e.g. [6]), obtained by linear smoothing of $\{(\mathbf{x}_i, Y(\mathbf{x}_i)) : i = 1, \ldots, n\}$, is the solution for α to the least squares minimization problem

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

where $\mathbf{H}$ is a $d \times d$ symmetric positive definite matrix; K is a d-dimensional kernel and $K_{\mathbf{H}}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$.
The trend estimates can be written as:

$$\hat{\boldsymbol{\mu}} = (\hat{\mu}(\mathbf{x}_1), \ldots, \hat{\mu}(\mathbf{x}_n))^t = \mathbf{S}\mathbf{Y},$$

where $\mathbf{S}$ is the *smoother matrix*.

- The local linear trend can be computed in practice with the `locpol` function of the `npsp` package [3].

Bandwidth selection

- The bandwidth matrix $\mathbf{H}$ controls the shape and size of the local neighborhood used to estimate $\mu(\mathbf{x})$.

- Traditional bandwidth selectors, such as cross validation (CV) or generalized cross validation (GCV), do not have a good performance for dependent data [6], since they tend to undersmooth the trend function.

- The modified cross-validation criteria (proposed in [1] for time series) can be adapted for spatial data, by minimizing:

$$MCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(Y(\mathbf{x}_i) - \hat{m}_{-N(i)}(\mathbf{x}_i)\right)^2, \tag{3}$$

where $\hat{m}_{-N(i)}(\mathbf{x}_i)$ denotes the trend estimate obtained when ignoring observations in a neighborhood $N(i)$ around $\mathbf{x}_i$ (in practice, the best results were obtained removing between 10% and 20% of the closest observations, depending on the strength of the spatial dependence). Note that the ordinary CV approach is a particular case with $N(i) = \{\mathbf{x}_i\}$.

- An alternative that takes the spatial dependence into account is suggested in [5], based on minimizing:

$$CGCV(\mathbf{H}) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)}{1 - \frac{1}{n\sigma^2}tr(\mathbf{S}\boldsymbol{\Sigma})}\right)^2, \tag{4}$$

where $\sigma^2 = C(\mathbf{0})$ is the variance (or sill) and $\boldsymbol{\Sigma}$ is the covariance matrix of the data.

- In practice, the bandwidth can be selected through the `h.cv` function of the `npsp` package.

Variogram estimation

- The natural approach to estimate the dependence consists in removing the trend and estimating the variogram from the residuals $\mathbf{r} = \mathbf{Y} - \mathbf{S}\mathbf{Y}$. The pilot local linear estimate $\hat{\gamma}(\mathbf{u})$ is the solution for α to the least squares minimization problem:

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

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

- Nevertheless, the residuals variability may be very different to that of the true errors:

$$Var(\mathbf{r}) = \boldsymbol{\Sigma} + \boldsymbol{S}\boldsymbol{\Sigma}\boldsymbol{S}^{\boldsymbol{t}} - \boldsymbol{\Sigma}\boldsymbol{S}^{\boldsymbol{t}} - \boldsymbol{S}\boldsymbol{\Sigma},$$

(see e.g. [2], Section 3.4.3, for the case of the linear trend estimator).

- As the bias due to the direct use of residuals in variogram estimation may have a significant impact on inference, a similar approach to that described in [4] will be considered. Using an iterative algorithm, the squared differences of the residuals are conveniently corrected and used to compute a bias-corrected pilot local linear variogram estimate.

- The bandwidth parameter G can be selected by minimizing the relative squared error:

$$\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\left(\frac{(r_i - r_j)^2}{2\hat{\gamma}_{-(i,j)}(\mathbf{x}_i - \mathbf{x}_j)} - 1\right)^2,$$

where $\hat{\gamma}_{-(i,j)}$ is obtained by excluding the pair (i, j).

- In practice, the local linear variogram estimates can be computed with functions `np.svar` and `np.svar.corr` of the `npsp` package. The generic function `h.cv` may be used to select the corresponding bandwidth.

- The final variogram estimate is obtained by fitting a "nonparametric" isotropic Shapiro-Botha variogram model [7], to the bias-corrected nonparametric pilot estimate (by using the function `fitsvar.sb.iso` of the `npsp` package).

Automatic modeling algorithm

- The poposed procedure for the joint estimation of the trend and the semivariogram is as follows:

1. Select an initial bandwidth to estimate the trend using the modified cross-validation criterion (3).
2. Compute the local linear trend estimate (2).
3. Obtain a bias-corrected pilot semivariogram estimate from the corresponding residuals (similar to that described in [4]).
4. Fit a valid Shapiro-Botha model [7] to the pilot semivariogram estimates.
5. Select a new bandwidth matrix for trend estimation with the CGCV criterion (4), using the fitted variogram model to approximate the data covariance matrix.
6. Repeat steps 2-4 until convergence.

- In practice, only two iterations are usually needed for this algorithm to converge.

Simulation results

- $N = 1,000$ samples were generated following model (1) on regular grids in the unit square of different sizes $n = 10 \times 10$, 15×15 and 20×20, with mean function:

$$\mu(x_1, x_2) = \sin(2\pi x_1) + 4(x_2 - 0.5)^2,$$

and random errors ε_i normally distributed with zero mean and isotropic exponential covariogram:

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

(for $\mathbf{u} \neq \mathbf{0}$), where c_0 is the nugget effect, c_1 is the partial sill ($c_1 = 1 - c_0$) and a is the practical range. The values considered were: $a = 0.3, 0.6$ and 0.9, $c_0 = 0, 0.2, 0.4$ and 0.8.

- The proposed procedure described above was use to compute estimates of the trend and variogram functions. To study the effect of the bias due to the direct use of residuals in variogram estimation, the uncorrected (residual) nonparametric semivariogram estimator was also used (in step 2). To study the effect of the bandwidth selection criteria in trend estimation, the results were compared to those using standard cross-validation (CV) and the optimal bandwidth $\mathbf{H}_{MASE}$, obtained by minimizing the mean averaged squared error:

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

where $\boldsymbol{\mu} = (\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))^t$ (note that $\mathbf{H}_{MASE}$ cannot be used in practice, due to its dependence on unknown terms).

- For the sake of brevity, only the results for the case of $n = 20 \times 20$, $a = 0.6$ and $c_0 = 0.2$ are reported here. A summary of the squared errors of the trend estimates is shown in Table 1, whereas Table 2 shows summaries of the relative squared errors of the semivariogram estimates. In general, a good performance of the proposed procedure (CGCV-Corrected) was observed in all simulation settings.

Table 1. Summary of squared errors of the trend estimates.

	MASE	CV	CGCV-Residual	CGCV-Corrected
mean	0.039	0.248	0.375	0.048
median	0.011	0.281	0.420	0.020
sd	0.109	0.135	0.216	0.073

Table 2. Summary of relative squared errors of the semivariogram estimates.

	MASE		CV		CGCV	
	Residual	Corrected	Residual	Corrected	Residual	Corrected
mean	0.097	0.046	0.307	0.210	0.234	0.058
median	0.096	0.017	0.342	0.229	0.249	0.029
sd	0.066	0.118	0.124	0.105	0.136	0.082

Application to real data

- The proposed methodology was applied to total precipitations (square-root of rainfall inches) during March 2016 recorded over 1053 locations on the continental part of USA. This data set is supplied with the `npsp` package for `R`. Figure 1 shows the observed values (a), the estimated trend function (b), the bias-corrected variogram estimates (c) and the corresponding kriging predictions (d).

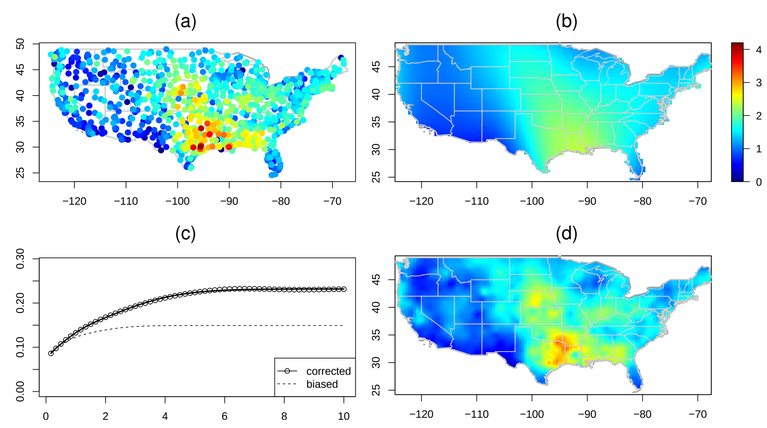

Figure 1. (a) Spatial locations and observed values, (b) nonparametric trend estimates, (c) semivariogram estimates, and (d) kriging predictions

Conclusions

- As observed in the simulation results, the proposed methodology seems to yield accurate estimates of the trend and variogram functions.

- The simulation results also lead to conclude that the bias, due to the direct use of residuals, may have a significant impact on variogram estimation (and consequently on the selected bandwidth with the CGCV criterion). Therefore, the use of the bias-corrected nonparametric variogram estimator would be recommended.

- Unlike traditional methods, as the approach is fully nonparametric, problems due to model misspecification are avoided.

- The proposed procedure was implemented in the function `np.fitgeo` of the `npsp` package [3] (available on CRAN).

Acknowledgments

This research has been supported by MINECO grant MTM2017-82724-R, and by the Xunta de Galicia (Grupos de Referencia Competitiva ED431C-2016-015 and Centro Singular de Investigación de Galicia ED431G/01), all of them through the European Regional Development Fund (ERDF).

References

[1] Chu, C. K. and Marron, J. S. (1991). Comparison of Two Bandwidth Selectors with Dependent Errors. *The Annals of Statistics* **19**, 1906–1918.

[2] Cressie, N. (1993). *Statistics for Spatial Data*. Wiley, New York.

[3] Fernández-Casal, R. (2019). npsp: Nonparametric Spatial Statistics. R package version 0.7-5. *http://github.com/rubenfcasal/npsp*.

[4] Fernández-Casal, R. and Francisco-Fernández, M. (2014). Nonparametric bias-corrected variogram estimation under non-constant trend. *Stochastic Environmental Research and Risk Assessment* **28**, 1247–1259.

[5] Francisco-Fernández, M. and Opsomer, J. D. (2005). Smoothing parameter selection methods for nonparametric regression with spatially correlated errors. *The Canadian Journal of Statistics* **33**, 279–295.

[6] Opsomer, J. D., Wang, Y. and Yang, Y. (2001). Nonparametric regression with correlated errors. *Statistical Science* **16**, 134–153.

[7] Shapiro, A. and Botha, J.D. (1991). Variogram fitting with a general class of conditionally non-negative definite functions. *Computational Statistics and Data Analysis* **11**, 87–96.