NONPARAMETRIC BOOTSTRAP APPROACH FOR RISK MAPPING UNDER HETEROSCEDASTICITY

S. Castillo-Páez[1], R. Fernández Casal[2], P. García Soidán[3]

[1]Universidad de las Fuerzas Armadas ESPE (Ecuador), [2]University of A Coruña (Spain), [3]University of Vigo (Spain); pgarcia@uvigo.es

Abstract

The aim of this work is to provide a nonparametric resampling method for approximating the (unconditional) probability that a spatial variable exceeds a prefixed threshold value. The existing approaches require assuming constant variance throughout the observation region, thus our proposal has been designed to be valid under heteroscedasticity of the spatial process. For this purpose, nonparametric estimates of the variance and variogram functions are computed by using corrected residuals, which are then employed to derive bootstrap replicates for approximation of the aforementioned risk. The performance of this mechanism is checked through numerical studies with simulated data.

Introduction

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

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

where μ and σ^2 denote the deterministic trend and variance functions, respectively, and ε is a second-order stationary process, with zero mean, unit variance and correlogram given by $\rho(\mathbf{u}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{u}))$. Then, the semivariogram of ε satisfies that $\gamma(\mathbf{u}) = \frac{1}{2}Var(\varepsilon(\mathbf{x}) - \varepsilon(\mathbf{x}+\mathbf{u})) = 1 - \rho(\mathbf{u})$.

- Suppose that n data $\mathbf{Y} = (Y(\mathbf{x}_1), \ldots, Y(\mathbf{x}_n))^t$ have been collected, at the respective locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

- The specification of the small-scale variability of the heteroscedastic process Y requires the estimation of σ^2 and ρ, since:

$$Cov(Y(\mathbf{x}), Y(\mathbf{x}+\mathbf{u})) = \sigma(\mathbf{x})\sigma(\mathbf{x}+\mathbf{u})\rho(\mathbf{u}).$$

Consequently, $\mathbf{\Sigma} = \mathbf{DRD}$, where $\mathbf{\Sigma}$ and $\mathbf{R}$ denote the covariance matrices of $\mathbf{Y}$ and $\boldsymbol{\varepsilon} = (\varepsilon(\mathbf{x}_1), \ldots, \varepsilon(\mathbf{x}_n))^t$, respectively, and $\mathbf{D} = diag(\sigma(\mathbf{x}_1), \ldots, \sigma(\mathbf{x}_n))$.

- Traditional methods for risk estimation focus on the approximation of $P(Y(\mathbf{x}_\alpha) > c | \mathbf{Y})$, namely, the conditional probability that $Y(\mathbf{x}_\alpha)$ exceeds a fixed threshold value $c > 0$. However, in some cases (e.g. climate studies), the aim is to study the distribution of the process under general conditions, thus asking for the estimation of the unconditional probability:

$$r_c(\mathbf{x}_\alpha) = P(Y(\mathbf{x}_\alpha) > c), \tag{2}$$

also called long-term risk [6].

- The current work is focused on the nonparametric estimation of $r_c(\mathbf{x}_\alpha)$ for heteroscedastic processes modeled as given in (1). To do the latter, a bootstrap technique is designed, based on the resampling approach derived in [4] under homoscedasticity, which requires approximating the functions μ, σ^2 and γ (or ρ).

Nonparametric modeling

- Firstly, a trend estimate is needed, which can be addressed through the local linear approach [2]:

$$\hat{\mu}(\mathbf{x}) = \mathbf{e}_1^t \left(\mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{X}_\mathbf{x}\right)^{-1} \mathbf{X}_\mathbf{x}^t \mathbf{W}_\mathbf{x} \mathbf{Y} = s_\mathbf{x}^t \mathbf{Y}, \tag{3}$$

where $\mathbf{e}_1 = (1, 0, \ldots, 0)$, $\mathbf{X}_\mathbf{x}$ is a matrix whose i-th row equals $(1, (\mathbf{x}_i - \mathbf{x})^t)$, $\mathbf{W}_\mathbf{x} = \mathtt{diag}\{K_\mathbf{H}(\mathbf{x}_1 - \mathbf{x}), ..., K_\mathbf{H}(\mathbf{x}_n - \mathbf{x})\}$, $K_\mathbf{H}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, K is a $d-$dimensional kernel function and $\mathbf{H}$ is the bandwidth matrix.

- Then, the estimates $\hat{\sigma}$ and $\hat{\gamma}$ are typically obtained from the residuals $\mathbf{r} = \mathbf{Y} - \mathbf{SY}$, where $\mathbf{S}$ is the *smoother matrix*, whose i-th row is equal to $\mathbf{s}_{\mathbf{x}_i}^t$.

- However, this mechanism tends to produce an underestimation of the small-scale variability of the process [1]. Indeed:

$$Var(\tilde{\boldsymbol{\varepsilon}}) = \mathbf{R} + \mathbf{B} = \mathbf{\Sigma}_{\tilde{\boldsymbol{\varepsilon}}}, \tag{4}$$

with $\tilde{\boldsymbol{\varepsilon}} = \mathbf{D}^{-1}\mathbf{r}$ denoting the standardized residuals and:

$$\mathbf{B} = \mathbf{D}^{-1}\left(\mathbf{S\Sigma S}^t - \mathbf{\Sigma S}^t - \mathbf{S\Sigma}\right)\mathbf{D}^{-1}. \tag{5}$$

- Relation (4) yields that:

$$Var\left(r_i/\sqrt{1+b_{ii}}\right) = \sigma^2(\mathbf{x}_i),$$
$$Var\left(\tilde{\varepsilon}(\mathbf{x}_i) - \tilde{\varepsilon}(\mathbf{x}_j)\right) = Var\left(\varepsilon(\mathbf{x}_i) - \varepsilon(\mathbf{x}_j)\right) + b_{ii} + b_{jj} - 2b_{ij}$$

where b_{ij} is the (i,j)-th element of the matrix $\mathbf{B}$ and $\tilde{\varepsilon}(\mathbf{x}_i) = r(\mathbf{x}_i)/\sigma(\mathbf{x}_i)$ is the i-th component of $\tilde{\boldsymbol{\varepsilon}}$.

- From these results, an iterative algorithm is designed for the joint estimation of σ^2 and γ, similar to that described in [3], although now the "exact" bias matrix (5) is used instead of an approximation to it. The specific steps are summarized below:

1. Use (3) to estimate the trend, compute the residuals $\mathbf{r}$ and obtain a pilot (uncorrected) estimate $\hat{\boldsymbol{\sigma}}^2 = (\hat{\sigma}^2(\mathbf{x}_1), \ldots, \hat{\sigma}^2(\mathbf{x}_n))$, by linear smoothing of $\left(\mathbf{x}_i, r_i^2\right)$.

2. Compute the estimated standardized residuals $\hat{\boldsymbol{\varepsilon}}_0 = \hat{\mathbf{D}}^{-1}\mathbf{r}$, where $\hat{\mathbf{D}} = diag(\hat{\sigma}(\mathbf{x}_1), \ldots, \hat{\sigma}(\mathbf{x}_n))$, and estimate the variogram $\gamma_{\hat{\varepsilon}}(u)$ by linear smoothing of $\left(||\mathbf{x}_i - \mathbf{x}_j||, (\hat{\varepsilon}_0(\mathbf{x}_i) - \hat{\varepsilon}_0(\mathbf{x}_j))^2\right)$.

3. Obtain a pilot estimation of the correlation matrix $\hat{\mathbf{R}} = \hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$ from $\hat{\gamma}_{\hat{\varepsilon}}(u)$.

4. Form $\hat{\mathbf{\Sigma}} = \hat{\mathbf{D}}\hat{\mathbf{R}}\hat{\mathbf{D}}$ and $\hat{\mathbf{B}} = \hat{\mathbf{D}}^{-1}\left(\mathbf{S}\hat{\mathbf{\Sigma}}\mathbf{S}^t - \hat{\mathbf{\Sigma}}\mathbf{S}^t - \mathbf{S}\hat{\mathbf{\Sigma}}\right)\hat{\mathbf{D}}^{-1}$.

5. Obtain an updated estimate $\hat{\boldsymbol{\sigma}}^2$ by linear smoothing of $\left(\mathbf{x}_i, r_i^2/(1+\hat{b}_{ii})\right)$.

6. Compute $\hat{\boldsymbol{\varepsilon}} = \hat{\mathbf{D}}^{-1}\mathbf{r}$ and approximate the error variogram by linear smoothing of $\left(||\mathbf{x}_i - \mathbf{x}_j||, (\hat{\varepsilon}(\mathbf{x}_i) - \hat{\varepsilon}(\mathbf{x}_j))^2 - \hat{b}_{ii} - \hat{b}_{jj} + 2\hat{b}_{ij}\right)$.

7. Obtain a new estimation of the correlation matrix $\hat{\mathbf{R}}$ and repeat steps 3-6 up to obtain convergence.

- We suggest selecting the bandwidth matrices from the corrected generalized cross-validation approach ($CGCV$), proposed in [5]: For instance, for the trend estimation this bandwidth is selected by minimizing:

$$CGCV(\mathbf{H}) = n^{-1}\sum_{i=1}^{n}\left(\frac{Y(\mathbf{x}_i) - \hat{\mu}_\mathbf{H}(\mathbf{x}_i)}{1 - n^{-1}tr\left(\mathbf{S}\hat{\mathbf{R}}\right)}\right)^2,$$

where $tr(\mathbf{A})$ denotes the trace of $\mathbf{A}$.

Bootstrap algorithm

- We propose the following algorithm to generate (unconditional) bootstrap replicas $Y^*(\mathbf{x}_\alpha)$ at the estimation locations $\{\mathbf{x}_\alpha : \alpha = 1, \ldots, n_0\}$, through a modification and extension of the proposal introduced in [4]:

1. Use the previous algorithm to approximate the trend and obtain the residuals (step 1), as well as the covariance matrix $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$ (step 3) and its Cholesky factorization $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}} = \mathbf{L}_{\hat{\varepsilon}}\mathbf{L}_{\hat{\varepsilon}}^t$.

2. Form the covariance matrix $\hat{\mathbf{\Sigma}}_\alpha$, corresponding to the estimation locations, by using the estimate $\hat{\gamma}$ derived in the previous algorithm (final step). Then, compute its Cholesky factorization $\hat{\mathbf{\Sigma}}_\alpha = \mathbf{L}_\alpha\mathbf{L}_\alpha^t$.

3. Compute the uncorrelated residuals $\mathbf{e} = (e_1, e_2, \ldots, e_n)^t = \mathbf{L}_{\hat{\varepsilon}}^{-1}\hat{\boldsymbol{\varepsilon}}_0$ and center them.

4. Obtain independent replicates of size n_0 from $\mathbf{e}$, denoted by $\mathbf{e}^* = \left(e_1^*, e_2^*, \ldots, e_{n_0}^*\right)^t$.

5. Compute the bootstrap errors $\varepsilon^* = (\varepsilon^*(\mathbf{x}_1), \ldots, \varepsilon^*(\mathbf{x}_{n_0}))^t = \mathbf{L}_\alpha\mathbf{e}^*$.

6. Use $Y^*(\mathbf{x}_\alpha) = \hat{\mu}_\mathbf{H}(\mathbf{x}_\alpha) + \hat{\sigma}(\mathbf{x}_\alpha)\varepsilon^*(\mathbf{x}_\alpha)$ to derive the final bootstrap sample.

- By repeating this scheme a large number of times, an estimate of the unconditional risk (2) at $\mathbf{x}_\alpha$ is provided by the proportion of values $\hat{Y}^*(\mathbf{x}_\alpha)$ exceeding the fixed threshold c.

Simulation results

- The new methodology was implemented in the statistical environment `R`, by using the nonparametric trend and variogram estimators supplied with the `npsp` package (available on CRAN).

- $1,000$ samples of different sizes n were generated from random processes Y following model (1) on a regular grid in the unit square, with trend function $\mu(x_1, x_2) = 2.5 + \sin(2\pi x_1) + 4(x_2 - 0.5)^2$ and two variance functions:
  - $\sigma_1^2(x_1, x_2) = 0.5(1 + x_1 + x_2)$
  - $\sigma_2^2(x_1, x_2) = \left(\frac{15}{16}\right)^2 \left(1 - (2x_1 - 1)^2\right)^2 \left(1 - (2x_1 - 1)^2\right)^2 + 0.1$

- The error process ε was assumed to be gaussian, with isotropic exponential variogram $\gamma_\theta(\mathbf{u}) = c_0 + c_1\left(1 - \exp\left(-3\frac{\|\mathbf{u}\|}{a}\right)\right)$, 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 in the simulations were: $n = 10 \times 10$, 15×15 and 20×20, $a = 0.3$, 0.6 and 0.9, and nugget values of $c_0 = 0$, 0.2, 0.4 and 0.8.

- Figures 1 and 2 compare the theoretical functions (variance and error variogram) and the average of their estimated counterparts through the nonparametric approach, showing the good performance of our proposal.

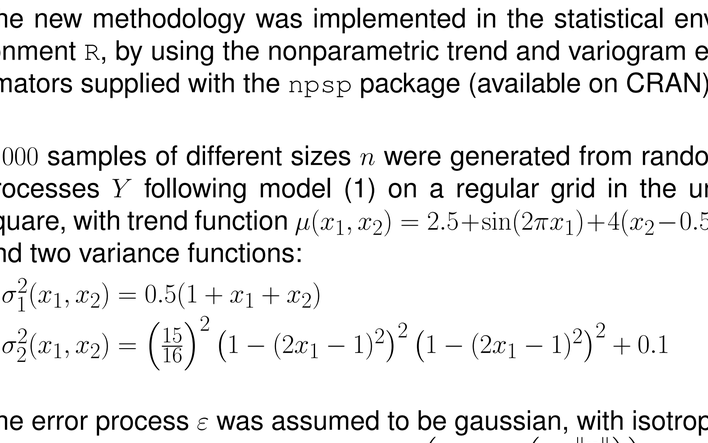

Figure 1. Theoretical variance (a) and average of the variance estimates (b), for σ_2^2, $n = 20 \times 20$, $r = 0.6$ and $c_0 = 0.2$.

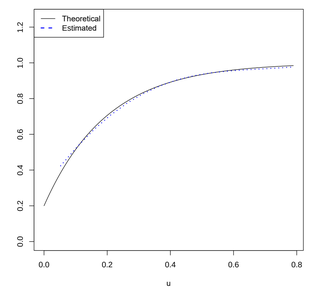

Figure 2. Theoretical error variogram (solid line) and average of the error variogram estimates (dotted line), for σ_2^2, $n = 20 \times 20$, $r = 0.6$ and $c_0 = 0.2$..

- The new Bootstrap approach was used to approximate $P(Y(\mathbf{x}_\alpha) > c)$ in a regular grid of $n_0 = 50 \times 50$ sites, with $c = 2.0, 2.5, 3.0, 3.5$ and 4.0. Figure 3 depicts the average of the resulting estimated probabilities for $c = 3.0$.

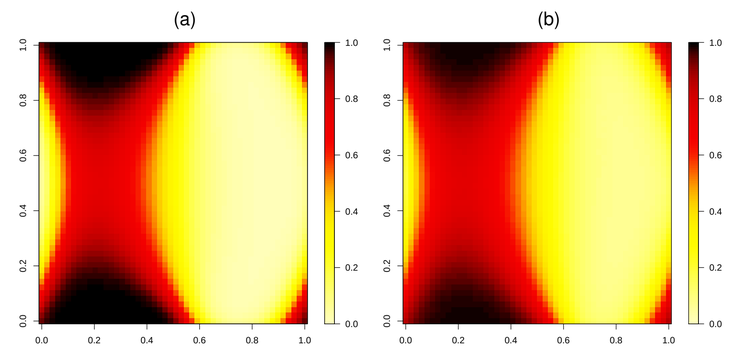

Figure 3. Maps of the theoretical (a) and averaged estimated (b) probabilities of exceeding $c = 3.0$, for σ_2^2, $n = 20 \times 20$, $r = 0.6$ and $c_0 = 0.2$.

- The squared errors are summarized in Table 1, giving account of the good behavior of the proposed procedure. Consistency of the new method is also observed, as the error statistics decrease for larger n.

Table 1. Summary of the squared errors ($\times 10^{-2}$) derived for the estimated probabilities, for σ_2^2, $r = 0.6$ and $c_0 = 0.2$..

	$n = 10 \times 10$			$n = 17 \times 17$			$n = 20 \times 20$		
c	mean	median	sd	mean	median	sd	mean	median	sd
2.0	1.84	0.08	4.41	1.77	0.07	4.16	1.71	0.07	4.04
2.5	2.75	0.47	5.58	2.72	0.45	5.49	2.63	0.44	5.35
3.0	2.59	0.55	5.44	2.53	0.53	5.29	2.46	0.52	5.27
3.5	2.08	0.19	4.74	2.06	0.18	4.67	2.00	0.18	4.60
4.0	1.44	0.02	4.16	1.45	0.02	4.22	1.41	0.02	4.13

- Table 2 shows the effect of spatial dependence on the risk estimation. As expected, when the range augments (higher dependence) the mean of the squared errors also increases. Similar conclusions remain valid when varying the nugget value.

Table 2. Averages of the squared errors ($\times 10^{-2}$) derived for the estimated probabilities under different spatial dependence settings, for σ_2^2, $n = 20 \times 20$, and $c_0 = 0.2$.

a	$c = 2.0$	$c = 2.5$	$c = 3.0$	$c = 3.5$	$c = 4.0$
0.30	1.17	1.82	1.58	1.25	0.94
0.60	1.71	2.63	2.46	2.00	1.41
0.90	2.06	3.13	2.98	2.44	1.67

Conclusions

- As observed in the simulation results, the proposed resampling method yields accurate estimates of the risk probabilities for a heteroscedastic spatial model.

- The new approach is fully nonparametric, thus avoiding the misspecification model problem.

- This methodology can be easily adapted to other issues, such as the construction of confidence and prediction intervals or the hypothesis testing under heteroscedasticity.

References

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

[2] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. Chapman & Hall, London.

[3] Fernández-Casal R., Castillo-Páez S. and García-Soidán P. (2017). Nonparametric estimation of the small-scale variability of heteroscedastic spatial processes. *Spatial Statistics* **22**, 358–370.

[4] Fernández-Casal R., Castillo-Páez S. and Francisco-Fernández M. (2018). Nonparametric geostatistical risk mapping. *Stochastic Environmental Research and Risk Assessment* **32**, 675–684.

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

[6] Krzysztofowicz, R. and Sigrest, A. A. (1997). Local climatic guidance for probabilistic quantitative precipitation forecasting. *Monthly Weather Review* **12**, 305-316.