Nonparametric Approximation of Conditional Risk
in Non-stationary Geostatistical Processes

R. Fernández Casal[1], S. Castillo-Páez[2], M. Francisco-Fernández[1]
[1] Universidade da Coruña (Spain), Centro de investigación CITIC; ruben.fcasal@udc.es, mariofr@udc.es
[2] Universidad de las Fuerzas Armadas ESPE (Ecuador); sacastillo@espe.edu.ec

Abstract

In this work, a nonparametric procedure to approximate the conditional probability that a regionalized variable exceeds a certain threshold value is proposed. The method consists of a bootstrap algorithm that combines conditional simulation techniques with nonparametric estimations of the trend and the variogram of the spatial process. For the local linear estimation of the mean, a bandwidth selection method that takes the spatial dependence into account is used. The variogram is approximated by a flexible estimator based on the residuals, previously correcting its bias due to the estimation of the trend. The proposed method allows obtaining estimates of the exceedance risk in non-observed spatial locations, and its behavior will be analyzed through simulation studies and with the application to a real data set.

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, using a fully nonparametric geostatistical approach, to estimate the conditional probability:

$$r_c(\mathbf{x}, \mathbf{Y}) = P(Y(\mathbf{x}) \geq c | \mathbf{Y})$$

where c is a threshold (critical) value.

- The geostatistical techniques commonly used to approximate this probability range from traditional methods, such as indicator kriging (e.g. [6]), to more recent procedures, such us those based on analysis of compositional data (e.g. [9]). However, these methods are designed for a constant trend and usually assume a parametric model for the variogram, 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, a general nonparametric procedure for spatial risk assessment is proposed. This procedure is a modification and an extension of the bootstrap method to estimate the unconditional probability $P(Y(\mathbf{x}) \geq c)$, proposed in [4].

Nonparametric estimation

- The local linear trend estimator (e.g. [7]), obtained by linear smoothing of $\{(\mathbf{x}_i, Y(\mathbf{x}_i)) : i = 1, \ldots, n\}$, can be written as:

$$\hat{\mu}_{\mathbf{H}}(\mathbf{x}) = \mathbf{S}\mathbf{Y},$$

where $\mathbf{S}$ is the *smoother matrix*, depending on a bandwidth matrix $\mathbf{H}$ that controls the shape and size of the local neighborhood used to estimate $\mu(\mathbf{x})$.

- The natural approach to estimate the dependence consists in removing the trend and estimating the variogram from the residuals $\mathbf{r} = \mathbf{Y} - \mathbf{S}\mathbf{Y}$. Nevertheless, the residuals variability may be very different to that of the true errors:

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

where $\boldsymbol{\Sigma}$ is the covariance matrix of the errors.

- As the bias due to the direct use of residuals in variogram estimation may have a significant impact on risk assessment, a similar approach to that described in [5] will be used. Using an iterative algorithm, the squared differences of the residuals are conveniently corrected and used to compute a pilot local linear variogram estimate. The final variogram estimate is obtained by fitting a "nonparametric" isotropic Shapiro-Botha variogram model [8], to the bias-corrected nonparametric pilot estimate.

Unconditional bootstrap

- We propose the following algorithm to generate unconditional bootstrap replicas $Y^*_{NS}(\mathbf{x}_\alpha)$ at the estimation locations $\{\mathbf{x}_\alpha : \alpha = 1, \ldots, n_0\}$ (modification of that described in Fernández-Casal *et al* [4])

1. Using the procedures described in previous section:
   (a) Compute $\hat{\mu}_{\mathbf{H}}(\mathbf{x})$ and the corresponding residuals $\mathbf{r}$ to obtain $\hat{\gamma}_{\mathbf{r}}(\cdot)$ and its corrected version $\hat{\gamma}(\cdot)$, following [5].
   (b) Form $\hat{\boldsymbol{\Sigma}}_{\mathbf{r}}$ from $\hat{\gamma}_{\mathbf{r}}(\cdot)$, and find the matrix $\mathbf{L}$ such that $\hat{\boldsymbol{\Sigma}}_{\mathbf{r}} = \mathbf{L}_{\mathbf{r}}\mathbf{L}_{\mathbf{r}}^t$, using Cholesky decomposition.
   (c) Form $\hat{\boldsymbol{\Sigma}}_\alpha$ corresponding to the estimation locations $\mathbf{x}_\alpha$ using $\hat{\gamma}(\cdot)$, and compute $\mathbf{L}_\alpha$ such that $\hat{\boldsymbol{\Sigma}}_\alpha = \mathbf{L}_\alpha\mathbf{L}_\alpha^t$.
2. Generate a bootstrap sample as follows:
   (a) Compute the "uncorrelated" residuals $\mathbf{e} = \mathbf{L}_{\mathbf{r}}^{-1}\mathbf{r}$ and center them.
   (b) Obtain independent bootstrap samples of size n_0 from $\mathbf{e}$, denoted by $\mathbf{e}^*$.
   (c) Compute the unconditional bootstrap errors $\varepsilon^*_{NS} = \mathbf{L}_\alpha\mathbf{e}^*$.
   (d) Obtain the unconditional bootstrap replicas $Y^*_{NS}(\mathbf{x}_\alpha) = \hat{\mu}_{\mathbf{H}}(\mathbf{x}_\alpha) + \varepsilon^*_{NS}(\mathbf{x}_\alpha),\ \alpha = 1, \ldots, n_0$.

- The latter algorithm uses unconditional simulation techniques based on Cholesky's decomposition. As the behavior of these replicas does not necessarily coincide with the observed values at the sample locations (see, e.g. [1], Section 7.3.1), this algorithm should not be used for conditional risk estimation.

Conditional bootstrap

- Taking into account the usual method to generate conditional simulations of stationary processes (combining unconditional simulation with kriging; see e.g. [1], Section 7.3.1), the proposed bootstrap algorithm to estimate the conditional risk is as follows:

1. Use the unconditional bootstrap algorithm described in previous section to (jointly) generate $\varepsilon^*_{NS}(\mathbf{x}_\alpha),\ \alpha = 1, \ldots, n_0$ and $\varepsilon^*_{NS}(\mathbf{x}_i),\ i = 1, \ldots, n$.
2. Compute the simple kriging predictions $\hat{\varepsilon}(\mathbf{x}_\alpha)$ and $\hat{\varepsilon}^*_{NS}(\mathbf{x}_\alpha)$ from the observed residuals $\mathbf{r}$ and from the bootstrap errors $\varepsilon^*_{NS}(\mathbf{x}_i)$, respectively.
3. Obtain the conditional bootstrap errors $\varepsilon^*_{CS}(\mathbf{x}_\alpha) = \hat{\varepsilon}(\mathbf{x}_\alpha) + \left[\varepsilon^*_{NS}(\mathbf{x}_\alpha) - \hat{\varepsilon}^*_{NS}(\mathbf{x}_\alpha)\right]$.
4. Compute the conditional bootstrap replicas $Y^*_{CS}(\mathbf{x}_\alpha) = \hat{\mu}_{\mathbf{H}}(\mathbf{x}_\alpha) + \varepsilon^*_{CS}(\mathbf{x}_\alpha)$.
5. Repeat steps 1 to 4 a large number of times B to obtain $Y^{*(1)}_{CS}(\mathbf{x}_\alpha), \ldots, Y^{*(B)}_{CS}(\mathbf{x}_\alpha)$.
6. Compute $\hat{r}_c(\mathbf{x}_\alpha, \mathbf{Y}) = \frac{1}{B}\sum_{j=1}^{B} I\left(Y^{*(j)}_{CS}(\mathbf{x}_\alpha) \geq c\right)$.

Simulation results

- Regular grids in the unit square of different sizes $n_1 = 16 \times 16$, 20×20 and 30×30 were considered. The top right diagonal was set as the estimation locations (see Figure 1 (a)) and the remaining ones as the sample ($n = n_1 - n_0$).

- $N = 1,000$ samples were generated following model (1) on the sample locations, with mean function $\mu(x_1, x_2) = 2.5 + \sin(2\pi x_1) + 4(x_2 - 0.5)^2$ (see Figure 1 (b)) 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.

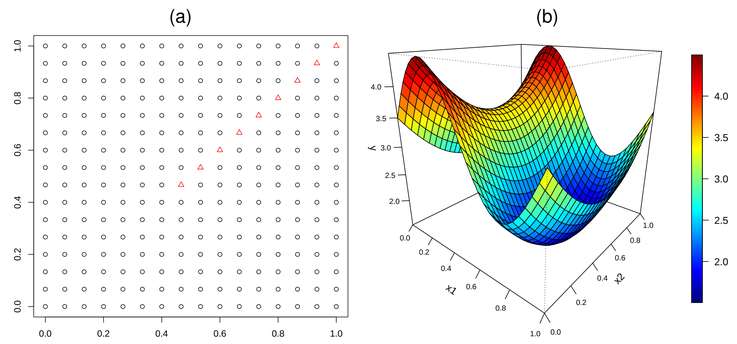

Figure 1. Sample and estimation locations (circles and triangles, respectively) for $n_1 = 16 \times 16$ (a) and theoretical trend (b).

- Using the proposed procedure, the conditional probabilities $P(Y(\mathbf{x}_\alpha) \geq c | \mathbf{Y})$ were estimated at each simulation, with $c = 2$, 3 and 4. Figure 2 shows the theoretical and estimated conditional risks for $c = 3$, $n_1 = 16 \times 16$, $a = 0.6$ and $c_0 = 0.2$.

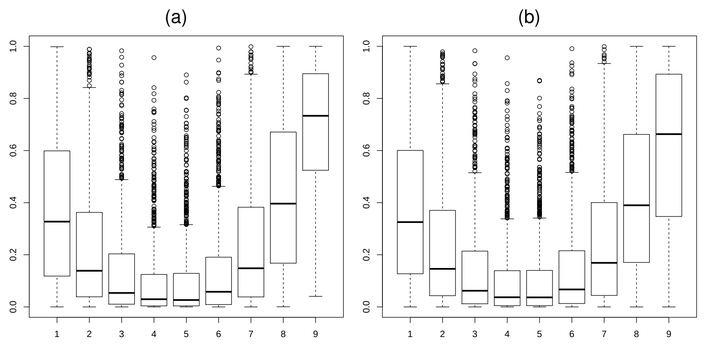

Figure 2. Box plots of the theoretical (a) and estimated (b) conditional probabilities of exceeding a threshold of 3.

- A summary of the squared errors ($\times 10^{-2}$), for $a = 0.6$ and $c_0 = 0.2$, is shown in Table 1. In general, a good performance of the proposed procedure was observed in all simulation settings.

Table 1. Summary of squared errors ($\times 10^{-2}$) of the estimated conditional probabilities.

	$n_1 = 16 \times 16$			$n_1 = 20 \times 20$			$n_1 = 30 \times 30$		
c	mean	median	sd	mean	median	sd	mean	median	sd
2	0.31	0.04	1.20	0.21	0.03	0.82	0.10	0.01	0.43
3	0.31	0.03	1.10	0.19	0.01	0.74	0.09	0.01	0.35
4	0.11	0.00	0.46	0.07	0.00	0.36	0.04	0.00	0.20

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 [3] for `R`. Figure 3 shows the observed values (a), the estimated trend function (b), the bias-corrected variogram estimates (c) and the kriging predictions (d).

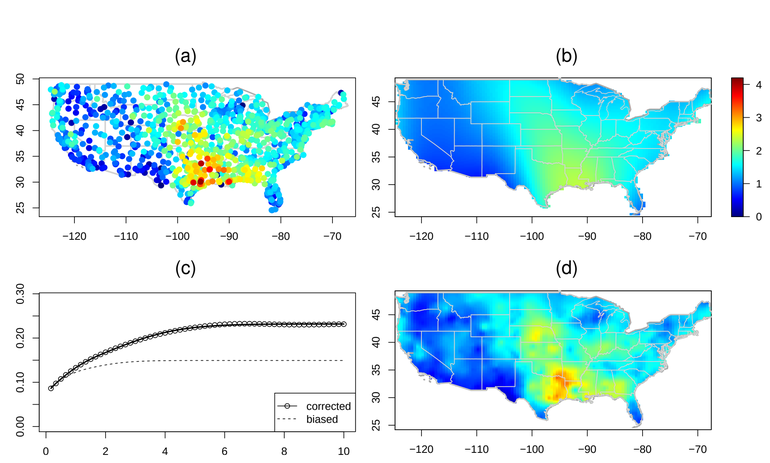

Figure 3. (a) Spatial locations and observed values, (b) nonparametric trend estimates, (c) semivariogram estimates, and (d) kriging predictions

- Applying the bootstrap algorithm described above, estimated probability maps for several critical values were computed. For instance, Figure 4 shows the estimated unconditional (a) and conditional (b) probabilities of occurring a total precipitation larger than or equal to the threshold $c = 2.0$ (square-root of rainfall inches).

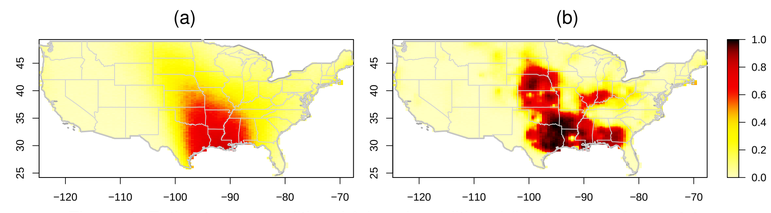

Figure 4. Estimated unconditional (a) and conditional (b) risk maps for $c = 2.0$.

Conclusions

- As observed in the simulation results, the proposed methodology yields accurate estimates of the conditional risk.

- Unlike traditional methods, as the approach is fully nonparametric, problems due to model misspecification are avoided. It can also be applied when the process exhibits a non constant trend.

- The procedure was implemented in the statistical environment `R`, using the functions for nonparametric trend and variogram estimation supplied with the `npsp` package [3] (available on CRAN).

Acknowledgments

The research of Rubén Fernández-Casal and Mario Francisco-Fernández has been supported by MINECO grants MTM2014-52876-R and 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 ERDF. The research of Sergio Castillo Páez has been supported by the Universidad de las Fuerzas Armadas ESPE.

References

[1] Chilès, J., and Delfiner, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*. Wiley & Sons, New York.

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

[3] Fernández-Casal, R. (2018). npsp: Nonparametric Spatial Statistics. R package version 0.7-2. *http://github.com/rubenfcasal/npsp*.

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

[5] 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.

[6] Goovaerts, P., Webster, R. and Dubois, P. (1997). Assessing the risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and Ecological Statistics* **4**, 31–48.

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

[8] 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.

[9] Tolosana-Delgado, R., Pawlowsky-Glahn, V., y Egozcue, J.-J. (2008). Indicator kriging without order relation violations. *Mathematical Geoscience*, **40**, 327–347