Nonparametric geostatistical risk mapping

R. Fernández-Casal[1,*], M. Francisco-Fernández[1], A. Quintela-del-Río[1]
and S. Castillo-Páez[2]

[1] *Facultad de Informática, University of A Coruña, 15071 A Coruña (Spain); ruben.fcasal@udc.es, mariofr@udc.es, aquintela@udc.es*
[2] *Facultad de CC Económicas y Empresariales, University of Vigo, 36310 Vigo (Spain); sacastillo@uvigo.es*
[*] *Corresponding author*

Abstract. *In this work, a fully nonparametric geostatistical approach to estimate threshold-exceeding probabilities is proposed. We suggest to use the nonparametric local linear regression estimator, with a bandwidth selected by a method that takes the spatial dependence into account, to estimate the large-scale variability (spatial trend) of a geostatistical process. To estimate the small-scale variability, a bias-corrected nonparametric estimate of the variogram is proposed. Finally, a bootstrap algorithm is used to estimate the probabilities of exceeding a threshold value at unsampled locations. The behavior of this approach is also evaluated through simulation and with an application to a real data set.*

Keywords. *Local linear regression; Nonparametric estimation; Bootstrap.*

1 Introduction

In spatial statistics, an effective method for spatial uncertainty assessment consist in estimating the probability or risk, given the data, that a spatial variable exceeds a threshold value. For instance, a map with estimated probabilities of pollutant concentrations exceeding a critical threshold can be very useful for environmental evaluation and decision making.

We assume that the spatial process $\left\{Y(\mathbf{x}), \mathbf{x} \in D \subset \mathbb{R}^d\right\}$ can be modeled as:

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

where $m(\cdot)$ is the trend function, accounting for the large-scale variability, and the error term ε, representing the small-scale variability, is a second order stationary process with zero mean and covariogram $C(\mathbf{h}) = Cov(\varepsilon(\mathbf{x}), \varepsilon(\mathbf{x}+\mathbf{h}))$.

In this framework, given n observed values $\{Y(\mathbf{x}_1),\ldots,Y(\mathbf{x}_n)\}$ of this process at locations $\{\mathbf{x}_1,\ldots,\mathbf{x}_n\}$, the goal is to make inference about the value of the variable Y at a location $\mathbf{x_0}$. Specifically, we are interested in estimating the conditional probability:

$$P\left(Y(\mathbf{x}_0) \geq c|\,\mathbf{Y}\right)$$

where c is a (critical) threshold value and $\mathbf{Y} = (Y(\mathbf{x}_1),...,Y(\mathbf{x}_n))^t$.

There are several geostatistical approaches which may be used to estimate the probabilities of exceeding a threshold value, such as indicator kriging (e.g. Goovaerts *et al.* [6]), disjunctive kriging (e.g. Webster and Oliver [10]) or Markov chain geostatistical modeling (Li *et al.* [7]), among others. However, the results obtained when these procedures are applied in practice could be unsatisfactory, usually due to the misspecification of the assumed parametric model (apart from other potential issues). 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.

2 Nonparametric geostatistical modeling

The local linear trend estimator (see, for instance, Opsomer *et al.* [8] and references therein) is given by:

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

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 $(1,(\mathbf{x}_i-\mathbf{x})^t)$, $\mathbf{W}_\mathbf{x} = \mathtt{diag}\{K_\mathbf{H}(\mathbf{x}_1-\mathbf{x}),...,K_\mathbf{H}(\mathbf{x}_n-\mathbf{x})\}$, $K_\mathbf{H}(\mathbf{u}) = |\mathbf{H}|^{-1} K(\mathbf{H}^{-1}\mathbf{u})$, K is a multivariate kernel and $\mathbf{H}$ is a $d \times d$ symmetric positive definite matrix.

The bandwidth matrix $\mathbf{H}$ controls the shape and size of the local neighborhood used for estimating $m(\mathbf{x})$. We recommend the use of the "bias corrected and estimated" generalized cross-validation (GCV) criterion proposed by Francisco-Fernández and Opsomer [4] to select this bandwidth in practice. This method consists in selecting the bandwidth $\mathbf{H}$ that minimizes:

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

being $\mathbf{S}$ the $n \times n$ matrix whose ith row is equal to $\mathbf{s}_{\mathbf{x}_i}^t$ (the smoother vector for $\mathbf{x} = \mathbf{x}_i$) and $\hat{\mathbf{R}}$ an estimate of the correlation matrix of the observations.

As in traditional geostatistical approaches, the usual dependence estimation method consists in removing the trend and estimating the variogram from the residuals:

$$\hat{\varepsilon} = \mathbf{Y} - \mathbf{S}\mathbf{Y}$$

Nevertheless, it is well-known that the direct use of the residuals in the estimation of the variogram (or the covariogram) may produce a strong underestimation of the small-scale variability of the process (e.g. Cressie [1], Section 3.4.3). Simply note that:

$$Var(\hat{\varepsilon}) = \mathbf{\Sigma} + \mathbf{S\Sigma S}^t - \mathbf{\Sigma S}^t - \mathbf{S\Sigma} = \mathbf{\Sigma}_{\hat{\varepsilon}},$$

where $\mathbf{\Sigma}$ is the covariance matrix of the errors. As this bias may also have a significant impact on the estimation of threshold-exceeding probabilities, a similar approach to that described in Fernández-Casal and Francisco-Fernández [2] will be used. The iterative approach starts with an initial nonparametric local linear fit of the trend. The squared differences of the corresponding 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 (Shapiro and Botha [9]), or an anisotropic extension (Fernández-Casal *et al.* [3]), to the bias-corrected nonparametric pilot estimate.

3 Bootstrap algorithm

The proposed bootstrap algorithm, designed to assess the variability of the fitted model, is a modification of the semiparametric bootstrap described in Francisco-Fernández *et al.* [5]. The specific steps are the following:

1. Using the procedures described in previous section:
   (a) Obtain the optimal bandwidth matrix $\mathbf{H}$, the residuals $\hat{\varepsilon}_i = Y(\mathbf{x}_i) - \hat{m}_{\mathbf{H}}(\mathbf{x}_i)$, $i = 1,\ldots,n$, and the estimated covariance function $\hat{C}$ of the errors, using the approach described above.
   (b) Using the estimated covariogram $\hat{C}$, compute the (estimated) covariance matrix of the errors $\hat{\mathbf{\Sigma}}$ and find the matrix $\mathbf{L}$, such that $\hat{\mathbf{\Sigma}} = \mathbf{L}\mathbf{L}^t$, using Cholesky decomposition.
   (c) Compute the (estimated) covariance matrix of the residuals $\hat{\varepsilon} = (\hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n)^t$, denoted by $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}}$. and the matrix $\mathbf{L}_{\hat{\varepsilon}}$, such that $\hat{\mathbf{\Sigma}}_{\hat{\varepsilon}} = \mathbf{L}_{\hat{\varepsilon}}\mathbf{L}^t_{\hat{\varepsilon}}$.
2. Generate a bootstrap sample with the estimated spatial trend $\hat{m}_{\mathbf{H}}(\mathbf{x}_i)$ and adding bootstrap errors generated as a spatially correlated set of errors. The bootstrap errors are obtained as follows:
   (a) Compute the "independent" variables $\mathbf{e} = (e_1, e_2, \ldots, e_n)^t$, given by $\mathbf{e} = \mathbf{L}_{\hat{\varepsilon}}^{-1}\hat{\varepsilon}$.
   (b) These independent variables are centered and, from them, we obtain an independent bootstrap sample of size n, denoted by $\mathbf{e}^* = (e_1^*, e_2^*, \ldots, e_n^*)^t$.
   (c) Finally, the bootstrap errors $\hat{\varepsilon}^* = (\hat{\varepsilon}_1^*, \ldots, \hat{\varepsilon}_n^*)^t$ are $\hat{\varepsilon}^* = \mathbf{L}\mathbf{e}^*$, and the bootstrap samples are $Y^*(\mathbf{x}_i) = \hat{m}_{\mathbf{H}}(\mathbf{x}_i) + \hat{\varepsilon}_i^*$, $i = 1, 2, \ldots, n$.
3. Compute the kriging prediction $\hat{Y}^*(\mathbf{x}_0)$ at each unsampled location $\mathbf{x}_0$ from the bootstrap sample (applying the nonparametric local linear regression estimator to the bootstrap sample, using the same bandwidth $\mathbf{H}$ as for the original analysis, and adding the simple kriging predictions obtained from the corresponding residuals).
4. Repeat steps 2 and 3 a large number of times B (in our analysis, $B = 1,000$). In a final point, a map with the frequencies (across bootstrap replicates) of how often a location is included in the at-risk area is computed.

Note also that this procedure can be adapted to the construction of confidence (prediction) intervals or hypothesis testing.

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

References

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

[2] 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.

[3] Fernández-Casal, R., González Manteiga, W. and Febrero-Bande, M. (2003). Flexible Spatio-Temporal Stationary Variogram Models. *Statistics and Computing* **13**, 127–136.

[4] 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.

[5] Francisco-Fernández, M., Quintela-del Río, A. and Fernández-Casal, R. (2011). Nonparametric methods for spatial regression. An application to seismic events. *Environmetrics* **23**, 85–93.

[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] Li, W. , Zhang, C., Dey, D. K. and Wang, S. (2010). Estimating threshold-exceeding probability maps of environmental variables with Markov chain random fields. *Stochastic Environmental Research and Risk Assessment* **24**, 1113–1126.

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

[9] 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.

[10] Webster, R. and Oliver, M.A. (1989). Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. *Journal of Soil Science* **40**, 497–512.