A kernel indicator variogram and its application to groundwater pollution data

Menezes, Raquel
University of Minho, Dept. of Mathematics and Applications
Campus de Azurém
4800-058 Guimarães, Portugal
E-mail: rmenezes@math.uminho.pt

Garcia-Soidán, Pilar
University of Vigo, Dept. of Statistics and Operations Research
Campus A Xunqueira
36005 Pontevedra, Spain
E-mail: pgarcia@uvigo.es

1. Introduction

An important problem in environmental sciences is the delineation of polluted zones with respect to regulatory standards. This can be done by approximating the distribution function of a spatial random process $\{Z(\mathrm{s}) : \mathrm{s} \in D \subset I\!R^d\}$, which provides the probability that the variable involved does not exceed a given threshold. For instance, suppose that one aims to assess soil contamination, then the approximation of the distribution function allows for the construction of probability risk maps.

Typically, a finite number of spatial locations s_i is selected, $1 \leq i \leq n$, where measurements of the variable involved are taken and used to derive information for the whole observation region, including the non-sampled locations. In this setting, estimation of the distribution function can be addressed in a parametric way, by imposing a shape or analytical expression for it. However, the data provided do not always support the distribution model assumption, so that a nonparametric approach must be adopted instead. In that case, the indicator kriging proves to be an efficient method (Journel, 1983), which provides an estimate of the distribution function at any spatial location.

The indicator approach is based on the interpretation of the distribution function as the expectation of an indicator random variable, namely:

$$\mathrm{P}\left(Z(\mathrm{s}) \leq x\right) = F_{\mathrm{s}}(x) = \mathrm{E}\left[I(\mathrm{s}, x)\right]$$

with $I(\mathrm{s}, x) = 1$ if $Z(\mathrm{s}) \leq x$ and zero otherwise. In practice, the distribution is approximated at Q previously fixed thresholds x_q and the remaining values are obtained by interpolation.

As mentioned in Goovaerts (1997), the least-squares (kriging) estimator of the indicator function is also the least-squares estimator of its expectation, according to the projection theorem. Then, an approximation of the distribution function at location s and threshold x is given by the indicator kriging predictor of $I(s, x)$, expressed as:

$$\hat{I}(\mathrm{s}, x) = \sum_{i=1}^{n} \lambda_i I(\mathrm{s}_i, x)$$

where $\{\lambda_i : 1 \leq i \leq n\}$, are obtained as the solution of the corresponding kriging equations. The latter implies that a variogram (or covariance function), which is known as indicator variogram (or indicator covariance function), needs to be inferred for each threshold. This work deals with this issue, which is called indicator structural analysis.

To develop this theory, we will assume that the random process is strictly stationary, so that $F_s(x) = F_{s'}(x) = F(x)$, for all $x \in I\!R$ and all $s, s' \in D$. Then, the indicator variogram is defined as:

$$2\gamma_I(\mathrm{t}, x) = \mathrm{Var}\left[I(\mathrm{s}, x) - I(\mathrm{s}+\mathrm{t}, x)\right] = \mathrm{E}\left[\left(I(\mathrm{s}, x) - I(\mathrm{s}+\mathrm{t}, x)\right)^2\right]$$

for each $t \in I\!R^d$ and $x \in I\!R$.

In this setting, an estimator of the indicator variogram is given by the experimental variogram, derived from the method of moments (Matheron, 1963):

$$2\hat{\gamma}_I(\mathrm{t}, x) = \frac{1}{N(\mathrm{t})} \sum_{(i,j)\in N(\mathrm{t})} \left(I(\mathrm{s}_i, x) - I(\mathrm{s}_j, x)\right)^2 \tag{1}$$

where $N(\mathrm{t})$ denotes the set of distinct pairs (i, j) satisfying that $\mathrm{s}_i - \mathrm{s}_j = \mathrm{t}$.

When data are irregularly spaced, the empirical estimator of the indicator variogram can be smoothed by considering instead a tolerance region $T(t)$ around t. The indicator kriging also demands the use of a valid variogram estimator, so that it fulfills the conditionally negative-definiteness property. We can proceed by first choosing a parametric family and then selecting that semivariogram in the family considered which best fit the data.

In the current work, we suggest a nonparametric alternative to the empirical variogram, which is similar to that analyzed in Garcia-Soidán (2007) and is adapted to the indicator setting:

$$2\hat{\gamma}_{I,h}(\mathrm{t}, x) = \frac{\sum_{i\neq j} K\left(\frac{\mathrm{t}-(\mathrm{s}_i-\mathrm{s}_j)}{h}\right)\left(I(\mathrm{s}_i, x) - I(\mathrm{s}_j, x)\right)^2}{\sum_{i\neq j} K\left(\frac{\mathrm{t}-(\mathrm{s}_i-\mathrm{s}_j)}{h}\right)} \tag{2}$$

where K represents a d-dimensional kernel function and h is the bandwidth parameter.

The above estimator considers a tolerance region around t, whose amplitude depends on h, and has the advantage that the more weight is given to each of the pairs considered, the closer to t is the lag between the locations involved. In consequence, the kernel indicator variogram provides a smoother estimator, whose consistency will be derived under several assumptions.

In addition, a direct estimation of the distribution function can be obtained through that of the sill of the indicator variogram, as proposed in Journel (1983). In this respect, the sill $S(x)$ of the indicator variogram is linked to the distribution function as follows:

$$S(x) = \lim_{\|\mathrm{t}\|\to\infty} \gamma_{I,h}(\mathrm{t}, x) = F(x) - F(x)^2$$

Then, approximation of the sill of the kernel indicator variogram provides another option for estimation of the distribution function.

2. Main results

2.1. Check consistency of (2)

Let $\{Z(\mathrm{s}) : \mathrm{s} \in D \subset I\!R^d\}$ be a spatial random process and denote by $Z(\mathrm{s}_1)$, ..., $Z(\mathrm{s}_n)$, the n data collected at the spatial locations s_1, ..., s_n.

We will assume that the random process is strictly stationary, namely:

(H1) $F_{\mathrm{s}_1,\ldots,\mathrm{s}_j}(x_1, ..., x_j) = F_{\mathrm{s}_1+d,\ldots,\mathrm{s}_j+d}(x_1, ..., x_j)$, for all $d \in I\!R^d$ and $j \geq 1$, with $F_{\mathrm{s}_1,\ldots,\mathrm{s}_j}(x_1, ..., x_j) = \mathrm{P}\left(Z(\mathrm{s}_1) \leq x_1, ..., Z(\mathrm{s}_j) \leq x_j\right)$

In particular, we will write F for the univariate distribution function, namely, $F_s = F$, for all $s \in I\!R^d$. In addition, we will ask:

(H2) For all $j \leq 4$ and $(\mathrm{s}_1, ..., \mathrm{s}_j) \in D^j$, $F_{\mathrm{s}_1,...,\mathrm{s}_j}(x_1, ..., x_j)$ admits two continuous derivatives in a neighborhood of $(\mathrm{s}_1, ..., \mathrm{s}_j)$, as a function of $(\mathrm{s}_1, ..., \mathrm{s}_j)$.

(H3) For all $(\mathrm{s}_1, \mathrm{s}_2) \in D^2$, $F_{\mathrm{s}_1,\mathrm{s}_2}(x_1, x_2)$ admits three continuous derivatives in a neighborhood of $(\mathrm{s}_1, \mathrm{s}_2)$, as a function of $(\mathrm{s}_1, \mathrm{s}_2)$.

The observation region D will be considered to be increasing and a random design will be assumed for the spatial locations, as suggested in Hall et al. (1994) to achieve consistent estimation:

(H4) $D = D_n = \beta D_0$, for some $\beta = \beta_n$ diverging to $+\infty$ and some bounded region $D_0 \subset I\!R^d$ containing a sphere with positive d-dimensional volume.

(H5) The spatial locations will be taken as $\mathrm{s}_i = \beta\, \mathrm{u}_i$, for $1 \leq i \leq n$, where u_1, ..., u_n represents a realization of a random sample of size n drawn from g_0, where g_0 is the density function considered on D_0.

(H6) For a given $\mathrm{t} \in I\!R^d$, g_0 admits three continuous derivatives in a neighborhood of t.

(H7) $\left\{h + (nh)^{-1} + \beta^{-1} + n^{-2}\beta^d h^{-d}\right\} \overset{n\to\infty}{\longrightarrow} 0$.

A dependence condition will be required from the random process, similar to that imposed in Zhu and Lahiri (2007) to establish a central limit theorem for the empirical process of a random field.

(H8) $\alpha(k, b) \leq c_1 k^{-c_2} b^{c_3}$, for some positive real numbers c_1, c_2, c_3.

(H9) K is a d-variate, compactly supported, symmetric and bounded density function, with $K(0) > 0$.

Under assumptions (H1)-(H9), estimator $\hat{\gamma}_{I,h}(\mathrm{t}, x)$ in (2) satisfies several properties, such as asymptotically unbiasedness and consistency, for all $\mathrm{t} \in I\!R^d$ and $x \in I\!R$. More specifically, we can check that:

$$\begin{aligned}
&\mathrm{Bias}\left[2\hat{\gamma}_{I,h}(\mathrm{t}, x)\right] = h^2 \sum_{k,l} \frac{\partial^2 \gamma_I(\mathrm{t}, x)}{\partial \mathrm{t}^{(k)} \partial \mathrm{t}^{(l)}} \int \mathrm{z}^{(k)} \mathrm{z}^{(l)} K(\mathrm{z}) d\mathrm{z} + o\left(h^2\right) \\
&\mathrm{Var}\left[2\hat{\gamma}_{I,h}(\mathrm{t}, x)\right] = 2n^{-2}\beta^d h^{-d} \gamma_I(\mathrm{t}, x) \left(\int g_0(\mathrm{u})^2 \, d\mathrm{u}\right)^{-1} \int K(\mathrm{z})^2 d\mathrm{z} + \\
&\quad + \beta^{-d} \left(\int g_0(\mathrm{u})^2 \, d\mathrm{u}\right)^{-2} \int g_0(\mathrm{u})^4 \, d\mathrm{u} \int \mathrm{E}\left[\left((I(0, x) - I(\mathrm{t}, x))^2 - 2\gamma_I(\mathrm{t}, x)\right) \cdot \right. \\
&\quad \left. \cdot \left((I(\mathrm{v}, x) - I(\mathrm{t} + \mathrm{v}, x))^2 - 2\gamma_I(\mathrm{t}, x)\right)\right] d\mathrm{v} + o\left(n^{-2}\beta^d h^{-d} + \beta^{-d}\right)
\end{aligned}$$

for all $\mathrm{t} \in I\!R^d$ and $x \in I\!R$.

In consequence, the MSE and the MISE of the kernel indicator variogram tend to zero as the sample size increases, so that minimization of the above quantities can provide asymptotically optimal bandwidth parameters. An alternative for selection of h may be that of considering a balloon estimator, namely, a kernel estimator where the bandwidth is allowed to vary with the lag t, as developed in Terrell and Scott (1992) for density estimation. For instance, we could take $h = h_k(\mathrm{t})$ as the euclidean distance from t to the k-nearest distances between locations in the sample.

2.2. Approximation of F(x)

The kernel indicator variogram can be used for approximation of the distribution function, either directly, as an application of the proposal given in Journel (1983), from now on denoted as the "Sill

Approach", or in an indirect way, by applying the "Kriging Approach". To proceed with the latter alternative, estimation of the distribution function F is typically discretized at a number of thresholds, previously fixed. Then, suppose that we select Q thresholds x_q. For each of them, the kernel indicator variogram $2\hat{\gamma}_{I,h}(.,x_q)$ must be obtained and used to solve the kriging equations, which provide Q values of the distribution function F. The remainder values can be approximated by interpolation.

The kernel indicator variogram can also be used for direct estimation of the distribution function. For this purpose, denote by $S(x)$ the sill of the kernel indicator variogram. Then, $S(x) = \hat{\gamma}_{I,h}(\infty, x) = F(x) - F(x)^2$. Now, take into account that $S(x)$ is increasing in $(-\infty, x_M]$ and decreasing in $[x_M, \infty)$ and takes values in [0, 0.25], where x_M stands for the median of the distribution F (Journel, 1983). In view of the latter, we can proceed as follows:

- Approximate the sill at each threshold, $\hat{S}(x_q)$.
- Determine the value of the median, $\hat{x}_M$, by selecting the value x_q for which $\hat{S}(x_q)$ is maximum and close to 0.25, so that $F(\hat{x}_M) \approx 0.5$.
- For each x_q, we can take $F(x_q) \approx 0.5\,(1 + \epsilon\,(x_q))\,\sqrt{1 - 4\hat{S}\,(x_q)}$, with $\epsilon(x) = sign\,(x - \hat{x}_M)$.

3. Simulation study

In order to analyse the performance of the proposed indicator semivariogram estimator, simulations of spatial data in IR^2 were carried out. Gaussian data were generated on the unit square $D \subset IR^2$, by selecting a theoretical variogram model to specify the spatial dependency. The new estimator given in (2) is compared against the estimator derived from the method of moments in (1). The symmetric Epanechnikov kernel was employed in the proposed kernel-type estimator. We considered a sample size $n = 50$ and two theoretical variograms, the exponential and the spherical, with a sill of $\sigma^2 = 1.5^2$, the corresponding range equal to $\phi = 0.3$ and no nugget effect $\tau^2 = 0$.

For a large number of datasets generated from a Gaussian spatial process, the main aims of the simulation study can be summarized as follows:

1. Compare two empirical $\hat{\gamma}_I(t,x)$ (Matheron and Kernel) with theoretical $\gamma_I(t,x)$ where
$$\gamma_I(t,x) = P\,[Z(s) \le x] - P\,[Z(s) \le x \cap Z(s+t) \le x] = F(x) - F_t(x,x);$$
2. Compare two valid $\tilde{\gamma}_I(t,x)$ (obtained from empirical $\hat{\gamma}_I$) with theoretical $\gamma_I(t,x)$;
3. Follow "Kriging Approach" and compare theoretical $F(x)$ with $\hat{F}(x) = \hat{I}(s,x) = \sum_{i=1}^{n} \lambda_i I(\mathrm{s}_i, x)$ where λ_i are kriging weights;
4. Follow 'Sill Approach" and compare theoretical $F(x)$ with $\hat{F}(x) = 0.5(1 + \hat{\epsilon}(x)\sqrt{1 - 4\hat{S}(x)}\,)$ where sill $\hat{S}(x)$ obtained from empirical $\hat{\gamma}_I$

Table 1 summarizes the results obtained when comparing the two empirical estimators (1) and (2). These results confirm that the proposed estimator offers a better approximation to the theoretical curve than the classic estimator obtained from the method of moments. This performance difference is quite evident, about 2.5 times, for the spherical model.

Table 2 summarizes the results of our second numerical study. When fitting a parametric model, two situations were considered, one assuming knowledge of the theoretical model used in simulation (referred to as Valid_LS1) and the other assuming misspecification (referred to as Valid_LS2). In the

Theoretical model	Variogram estimator	ISE values Mean (St. dev.)
Exponential	Matheron	3.93 (2.46)
	Kernel	2.66 (2.57)
Spherical	Matheron	2.45 (1.00)
	Kernel	0.98 (0.83)

Table 1: ISE values obtained for the Matheron and the kernel estimators, given in (1) and (2), respectively. All values were multiplied by 10^3.

latter case, if data are simulated with the spherical variogram, then the fitted model is the exponential one, and vice-versa.

The results of Table 2 show that, if the theoretical model is known, the best approach is to adopt the proposed kernel estimator and then fit this to the known parametric model through OLS. Otherwise, fitting should follow the method proposed by Shapiro and Botha (1991) (referred to as Valid_SB), after also adopting the kernel estimator. With respect to the standard deviations of ISE, the lower values are always associated to the Shapiro and Botha method.

Theoretical model	Variogram estimator	ISE values		
		Valid_LS1 Mean (St. dev.)	Valid_LS2 Mean (St. dev.)	Valid_SB Mean (St. dev.)
Exponential	Matheron	2.34 (2.77)	2.73 (4.44)	2.88 (2.71)
	Kernel	2.22 (2.69)	2.60 (4.39)	2.43 (2.59)
Spherical	Matheron	0.72 (0.90)	1.48 (1.65)	0.92 (0.58)
	Kernel	0.64 (0.85)	1.32 (1.11)	0.64 (0.50)

Table 2: ISE values obtained for the valid versions of the Matheron and the kernel estimators, given in (1) and (2), respectively. All values were multiplied by 10^3.

The following simulation study was planned to approximate the distribution function $F_{\mathrm{s}}(x)$, with $\mathrm{s} = (0.5, 0.5)$, by applying the two approaches described in Section 2.2., which were referred to as the "Kriging Approach" or the "Sill Approach".

As in the previous simulations studies, we here have compared both proposals under two scenarios, when adopting the proposed kernel estimator (2) and when adopting the Matheron estimator (1). Three thresholds were selected, so that $Q = 3$, identifying the quartiles 25%, 50% and 75% as being representative of the distribution domain, which will be respectively denoted by Q_1, Q_2 and Q_3. Results are summarized in Table 3, through the values of the mean and the standard deviations derived for the MSE.

From the first two lines of Table 3, we observe that the MSE values obtained for approximating $F(x)$ by sill estimation are also smaller under the option of kernel-type estimation. Furthermore, these same results can compete with the values of MSE associated to the indicator kriging approach (regardless of the estimator option), with the great advantage of avoiding to solve the kriging system. Consequently, one might conclude that approximation of the sill of the kernel indicator variogram offers an important mechanism for estimation of the distribution function.

Distribution approximation	Variogram estimator	MSE values		
		Q_1 Mean (St. dev.)	Q_2 Mean (St. dev.)	Q_3 Mean (St. dev.)
Sill	Matheron	0.227 (0.138)	0.104 (0.111)	0.091 (0.146)
	Kernel	0.196 (0.097)	0.012 (0.018)	0.036 (0.048)
IK	Matheron	0.096 (0.151)	0.095 (0.108)	0.087 (0.150)
	Kernel	0.082 (0.115)	0.097 (0.107)	0.082 (0.149)

Table 3: MSE values obtained for the two approximations of the distribution function, based on using the sill estimation and the indicator kriging. These approximations are based on Matheron and kernel estimators, given in (1) and (2), respectively. Data simulated with the spherical model. Total number of replicas is 100 and each sample size is 50.

4. Application to environmental data

We now present a practical situation where the issue of interest is to approximate the distribution function of a spatially distributed measure of groundwater quality, as it provides the probability that the associated random process $Z(\mathrm{s})$ does not exceed a given threshold.

Our example is related to groundwater quality data collected in Beja district (in the south of Portugal) in 1998 and 2000. The observation region $D \subset I\!R^2$ forms an approximated 50km^2 square area, as illustrated in the top panels of Figure 1. Measurements of nitrate were taken, as this chemical element is quite related with the agricultural activity, which is of great importance in this area. On the other hand, Beja district is part of one of the driest regions of Portugal, thus making the quality of water a very important issue.

Groundwater quality is regulated by the European Rule 2006/118/CE, concerning protection against pollution, and by the Portuguese Law n^{o}306/2007, related to the human uses. The maximum admissible value of the nitrate concentration for which the water is classified as potable is 50mgNO3/L. Aiming to determine the probability that the concentration of this pollutant does not exceed this critical value, at a certain unsampled location, we have considered the indicator variable $I(\mathrm{s}, x)$ with x =50mgNO3/L and $Z(\mathrm{s})$ measuring the nitrate concentration at location s.

The top panels of Figure 1 show the locations of the monitoring stations for each year. The size of each bullet is proportional to the indicator value found in each location, so that large bullets identify values under the threshold 50 mgNO3/L. The bottom panels display the approximations obtained for the indicator semivariograms based on Matheron and kernel estimators, defined in (1) and (2), respectively. Paralta and Ribeiro (2003) present kriging maps for the same indicator variables, after fitting a spherical model to the Matheron's estimates. Therefore, in our study, we have also followed these authors' approach aiming to compare our results with theirs. Furthermore, we have derived a valid version of our kernel proposal through Shapiro and Botha method avoiding some possible misspecification.

We now aim to construct pollution risk maps with these data. In particular, we focused on the estimation of the probability that the nitrate concentration did not exceed the maximum value admitted for human consumption along the Beja district, namely, $F(50)$. With this aim, we applied the two approaches, based either on the sill estimation or on the indicator kriging method, by adopting the proposed kernel estimator.

Figure 2 displays the results obtained over a grid of a total of 500 points, allowing us to identify the location of high risk areas. Note that higher probabilities of not exceeding the threshold 50mgNO3/L are present in 1998 than in 2000, which is consistent with an overall increment in levels

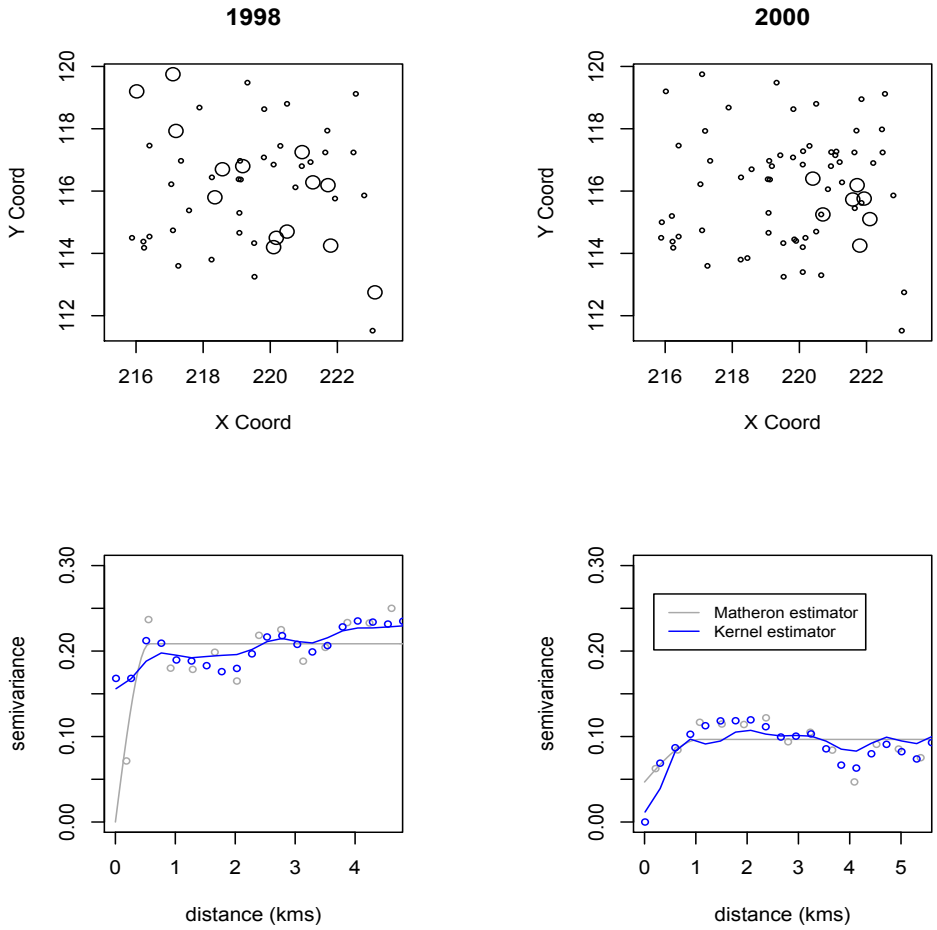

Figure 1: Top panels give the sample locations for 1998 and 2000 in Beja data. Bottom panels present the empirical estimators, given in (1) and (2), and the corresponding valid versions (full lines) for nitrate concentration data in Beja district. The distance unit is 1km.

of pollution over the two years between both surveys. Furthermore, one should notice that the two approaches for estimation of $F(50)$ offer similar outputs, although the approximation based on the indicator kriging approach seems smoother than that based on the sill estimation. This can be explained because the latter one involves local information.

REFERENCES (RÉFERENCES)

García-Soidán P. 2007. Asymptotic normality of the Nadaraya-Watson semivariogram estimator. *TEST* **16**(3): 479-503. DOI: 10.1007/s11749-006-0016-8.

Goovaerts P. 1997. *Geostatistics for natural resources evaluation* (1st ed.), vol. **1**: 284-330. Oxford University Press: New York.

Hall P, Patil P. 1994. Properties of nonparametric estimators of autocovariance for stationary random fields. *Probability Theory and Related Fields* **99**(3): 399-424. DOI: 10.1007/BF01199899.

Journel AG. 1983. Nonparametric estimation of spatial distribution. *Mathematical Geology* **15**(3): 445-468. DOI: 10.1007/BF01031292.

Matheron G. 1963. Principles of Geostatistics. *Economic Geology* **58** (8): 1246-1266. DOI: 10.2113/gsecongeo.58.8.1246.

Paralta E, Ribeiro L. 2003. Monitorização e Modelação Estocástica da Contaminação por Nitratos do Aquífero

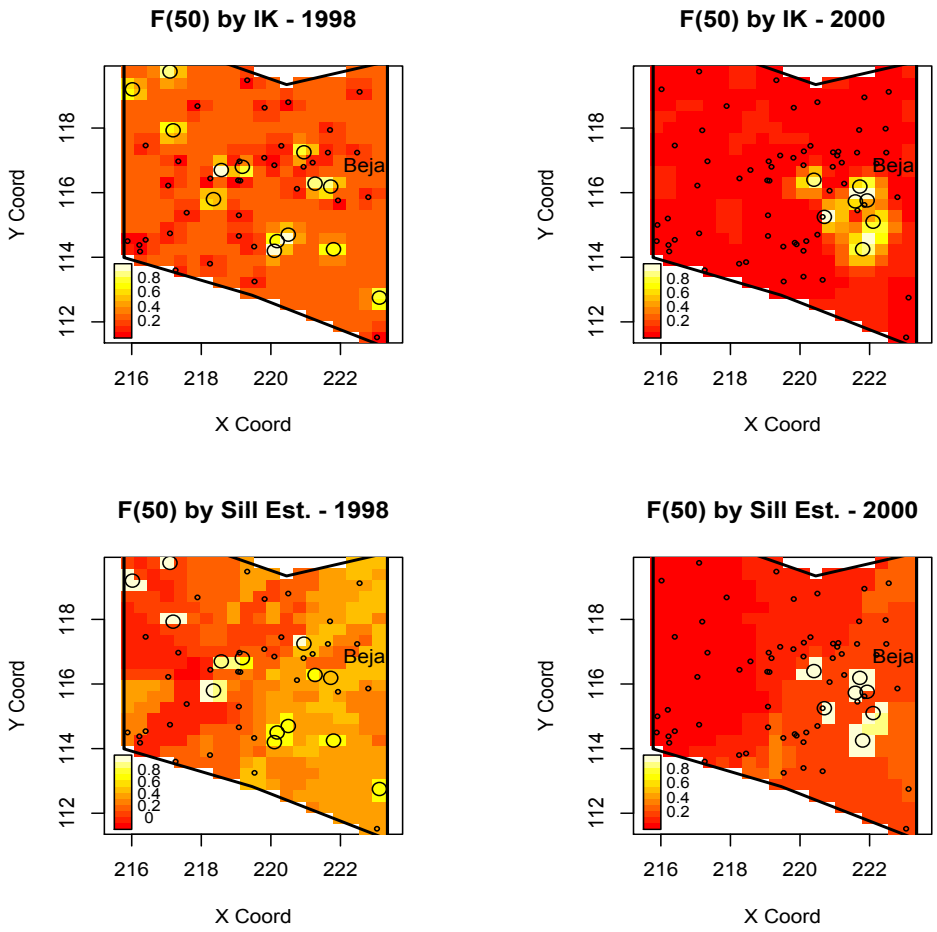

Figure 2: Pollution risk maps for nitrate concentration data in Beja district, displaying the estimates of $F(50)$, that is the probability of not exceeding the threshold $x = 50$, based on using indicator kriging and sill estimation.

Gabro-diorítico na Região de Beja. Resultados, Conclusões e Recomendações. Seminário sobre Águas Subterráneas, LNEC, Lisboa.

Shapiro A, Botha J. 1991. Variogram fitting with a general class of conditionally nonnegative definite functions. *Comput. Stat. Data Anal.* **11** (1): 87-96. DOI: 10.1016/0167-9473(91)90055-7.

Terrell G, Scott DW. 1992. Variable Kernel Density Estimation. *Annals of Statistics* **20**(3): 1236-1265.

Zhu J, Lahiri SN. 2007. Bootstrapping the Empirical Distribution Function of a Spatial Process. *Statistical Inference for Stochastic Processes* **10** (2): 107-145. DOI: 10.1007/s11203-005-2349-4.