Journal of Nonparametric Statistics
Vol. 00, No. 00, 00 00, 1–12

Spatial prediction via the kernel method with cross-validation approaches for bandwidth selection

Raquel Menezes[a]*, Célia Ferreira[a] and Pilar García-Soidán[b]

[a] *Department of Mathematics for Science and Technology, University of Minho, Portugal*;
[b] *Department of Statistics and Operations Research, University of Vigo, Spain*

(*00 released 00*)

In this work, the nonparametric kernel prediction will be considered for stochastic processes, when a random design is assumed for the spatial locations. We will check that, under rather general conditions, the mean-squared prediction error tends to be negligible, as the sample size increases. However, the use of the optimal bandwidth demands the estimation of unknown quantities, whose approximation in an accurate way often turns out to be difficult in practice. Hence, alternative cross-validation approaches will be provided for the selection of both local and global bandwidths. Numerical studies were carried out in order to analyse the performance of the nonparametric predictor for both simulated and real data.

Keywords: kernel method; prediction; stationarity.

AMS Subject Classification: 62G05; 62G10.

1. Introduction

A fundamental problem in spatial statistics is that of reconstructing a phenomenon over its domain from a discrete set of observed values. The kriging techniques are typically used for the latter purpose, providing us with predictors that are optimal in some sense. In fact, the referred approaches are derived by minimizing the mean-squared prediction error, subject to some constraints that are dependent on the hypotheses assumed from the random process.

Stationarity is a typical requirement and, under this condition, linear techniques have been developed to provide us with spatial predictors (e.g. [1, 2]). However, the results of the kriging equations rely on the validity of the conditions required, so that a failure in the hypotheses may have a significant effect. For instance, the misspecification of the distribution, of the mean or of the second-order structure may lead to poor predictions.

Taking the above in mind, we propose an alternative that may be obtained via a nonparametric approach. On this occasion, the kernel method has been considered, as it has been extensively used in the statistics literature, due to its simplicity and applicability to a wide range of problems, such as problems on density or regression estimation (see e.g. [11, 13, 14]).

In the spatial setting, kernel approaches have been suggested for the estimation of the dependence structure, in terms of the covariance function (e.g. [8]) or of the semivariogram (e.g. [6]). The aim of this work is to introduce a nonparametric

*Corresponding author. Email: rmenezes@mct.uminho.pt

ISSN: 1048-5252 print/ISSN 1029-0311 online

DOI: 10.1080/1048525YYxxxxxxxx
http://www.informaworld.com

kernel predictor, which proves to be valid under rather general conditions. In particular, we will check that the mean-squared prediction error tends to be negligible, as the sample size increases.

The implementation of the kernel method requires specification of a bandwidth parameter. Although it might be chosen subjectively by eye, the selection of the bandwidth is typically recommended to be derived from data. Possible options, such as that involving sample splitting, are described in [7]. The most common approach is based on the minimization of the corresponding error, providing us with a consistent bandwidth, and use plug-in estimates for approximation of the unknown terms [9, 12].

Proceeding in this way, we have derived the optimal bandwidth for kernel prediction, which demands the estimation of quantities dependent on the first and on the second order moments of the random process. Approximation of the referred terms in an accurate way often turns out to be difficult. Hence, cross-validation approaches are more easily attainable for a given data set, as explained in [10] and references therein, which are also suggested in this work for the selection of the bandwidth.

Finally, we will describe some numerical studies carried out in order to analyse the performance of the nonparametric predictor, when adopting different selections of the bandwidth, which will be compared with the results achieved by kriging predictors, for gaussian and non-gaussian data. An application of the proposed predictor to a real data set is also included.

This paper is organized as follows. Section 2 introduces the main hypotheses to be assumed. In section 3, the kernel prediction is developed, where the dominant terms of the mean squared prediction error are established and bandwidth selection is discussed. The simulation studies and the application to real data are detailed in sections 4 and 5, respectively.

2. Hypotheses

In this section, we will introduce the main conditions to be required so as to guarantee the validity of our results. Firstly, the hypotheses assumed from the spatial random process $Z(x)$ will be detailed, focussed on the type of stationarity imposed.

(H1) $\{Z(x) : x \in D \subset \mathbb{R}^d\}$ is a second-order stationary process with covariance function C, namely that it satisfies the following conditions:
  (i) $\mathrm{E}[Z(x)] = \mu$, for all $x \in D$ and some $\mu \in \mathbb{R}$.
  (ii) $\mathrm{Cov}[Z(x), Z(y)] = C(x - y)$, for all $x, y \in D$ and some function C.

(H2) C is three-times continuously differentiable in a neighbourhood of 0.

(H3) D is a bounded region with positive d-dimensional volume.

As regards the spatial locations, a random design will be assumed in order to achieve consistent estimation. Then, let f be a density function considered on D. We will denote by $(X_1, ..., X_n)$ a random sample of size n drawn from f and by $(x_1, ..., x_n)$ a realization of it. The density will be required to satisfy:

(H4) f is bounded and $f(x) > 0$ for all $x \in D$.

(H5) f is three-times continuously differentiable in a neighbourhood of x, for all $x \in D$.

The following hypotheses will be related to the kernel function as well as to the convergence rates required from the bandwidth parameter h.

(H6) K_d is a d-variate, compactly supported, symmetric and bounded density function, satisfying that $K_d(0) > 0$.

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

3. Main results

As remarked in the previous section, we will address our attention to the second-order stationary processes. Let $\{Z(x) : x \in D \subset I\!R^d\}$ be a random process, with constant mean μ and covariance function C.

Suppose that n data, $\mathbf{Z} = \{Z(x_1), ..., Z(x_n))\}$, are collected, at known spatial locations x_1, ..., x_n. If our aim is that of predicting the value of the random process at any given $x \in D$, we suggest to use the following nonparametric predictor:

$$\hat{Z}(x;h) = \frac{\sum_{i=1}^{n} w_i Z(x_i)}{\sum_{i=1}^{n} w_i}$$

which will be referred to as predictor NP, where $w_i = K_d\left(\frac{x-x_i}{h}\right)$, K_d represents a d-dimensional kernel function and h is the bandwidth parameter.

One advantage when applying the above predictor is that implementation of $\hat{Z}(x;h)$ does not require estimation of the second order structure of the random process, unlike that of kriging approaches.

Next, we will establish the dominant terms in the mean-squared prediction error of the NP predictor.

Theorem 3.1: *Assume that* $\{Z(x) : \mathrm{x} \in D \subset I\!R^d\}$ *is a stationary random process and that conditions H1-H7 are satisfied. Then, for* $x \in D$*, one has:*

$$\mathrm{E}\left[\left(Z(x) - \hat{Z}(x;h)\right)^2\right] = 2n^{-1}h^{2-d}f(x)^{-2}\sum_{i=1}^{d}\sum_{j=1}^{d} m_{2,i,j}A_{1,i,j}(x) -$$

$$-4h^2 f(x)^{-2}\sum_{i=1}^{d}\sum_{j=1}^{d} m_{1,i,j}A_{2,i,j}(x) + o\left(n^{-1}h^{2-d} + h^2\right)$$

where $A_{1,i,j}(x) = \frac{\partial C}{\partial z^{(i)}}\Big|_0 \frac{\partial f}{\partial z^{(j)}}\Big|_x - \frac{f(x)}{2}\frac{\partial^2 C}{\partial z^{(i)}\partial z^{(j)}}\Big|_0$, $A_{2,i,j}(x) = \frac{\partial C}{\partial z^{(i)}}\Big|_0 \frac{\partial f}{\partial z^{(j)}}\Big|_x$ *and* $m_{p,i,j} = \int z^{(i)}z^{(j)}K_d(z)^p dz$ *with* $p = 1, 2$.

Proof: Take into account that the stochastic process is second-order stationary over the observation region D, under H1, and that the spatial locations have been assumed to be generated at random from a density f on D. Then, one has:

$$\mathrm{E}[(Z(x) - \hat{Z}(x;h))^2] = \mathrm{E}[\mathrm{E}[(Z(x) - \hat{Z}(x;h))^2/X_1, ..., X_n]] =$$

$$= 2\mathrm{E}\left[\frac{a_2(x) + a_3(x)}{a_1(x)^2}\right] \tag{1}$$

where:

$$\begin{array}{l} a_1(x) = \sum_{k=1}^{n} W_k \\ a_2(x) = \sum_{k=1}^{n} W_k^2 \left(C(0) - C\left(x - X_k\right)\right), \\ a_3(x) = \sum_{k \neq l} W_k W_l \left(C\left(X_l - X_k\right) - C\left(x - X_k\right) - C\left(x - X_l\right) + C(0)\right) \end{array}$$

with $W_k = K_d\left(\frac{x-X_k}{h}\right)$.

At the end of this proof, we will state the following orders for $x \in D$:

$$\begin{array}{l} \alpha_1(x) = \mathrm{E}\left[K_d\left(\frac{x-X_k}{h}\right)/X_1, ..., X_n\right] = h^d f(x) + o\left(h^d\right) \\ \alpha_2(x) = \mathrm{E}\left[K_d\left(\frac{x-X_k}{h}\right)^2 \left(C(0) - C\left(x - X_k\right)\right)/X_1, ..., X_n\right] = \\ \quad = h^{d+2} \sum_{i=1}^{d} \sum_{j=1}^{d} m_{2,i,j} A_{1,i,j}(x) + o\left(h^{d+2}\right) \\ \alpha_3(x) = \mathrm{E}\left[K_d\left(\frac{x-X_k}{h}\right)\left(C\left(X_l - X_k\right) - C\left(x - X_k\right) - C\left(x - X_l\right) + C(0)\right)/X_1, ..., X_n\right] = \\ \quad = -2h^{2d+2} \sum_{i=1}^{d} \sum_{j=1}^{d} m_{1,i,j} A_{2,i,j}(x) + o\left(h^{2d+2}\right) \end{array}$$

Now, by applying similar arguments as those used in the proofs of Theorems 3.1 and 3.2 in [6] and conditions H3-H4, the following orders hold:

$$\begin{array}{l} a_1(x) = nh^d f(x) + o\left(nh^d\right) \ \ a.s., \\ a_2(x) = nh^{d+2} \sum_{i=1}^{d} \sum_{j=1}^{d} m_{2,i,j} A_{1,i,j}(x) + o\left(nh^{d+2}\right) \ \ a.s., \\ a_3(x) = -2n^2 h^{2d+2} \sum_{i=1}^{d} \sum_{j=1}^{d} m_{1,i,j} A_{2,i,j}(x) + o\left(n^2 h^{2d+2}\right) \ \ a.s. \end{array}$$

The above relations together with (1) would allow to conclude the validity of Theorem 3.1.

Then, it only remains to check the orders established for $\alpha_i(x)$, $i = 1, 2, 3$. To do the latter, we will bear in mind that:

$$\alpha_1(x) = \int K_d\left(\tfrac{x-u}{h}\right) f(u)\, du = h^d \int K_d(z)\, f(x - hz)\, dz = h^d f(x) + o\left(h^d\right)$$

by using the change of variable $z = \frac{x-u}{h}$, the fact that f is the density of the spatial locations as well as hypotheses H5-H7.

On the other hand:

$$\begin{array}{l} \alpha_2(x) = \int K_d\left(\frac{x-u}{h}\right)^2 \left(C(0) - C\left(x - u\right)\right) f(u)\, du = \\ \quad = h^d \int K_d(z)^2 \left(C(0) - C\left(hz\right)\right) f(x - hz)\, dz = \\ \quad = h^d \int K_d(z)^2 \cdot \\ \qquad \cdot \left(-h \sum_{i=1}^{d} z^{(i)} \left.\frac{\partial C}{\partial z^{(i)}}\right|_0 - \frac{h^2}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} z^{(i)} z^{(j)} \left.\frac{\partial^2 C}{\partial z^{(i)} \partial z^{(j)}}\right|_0 + ...\right) \cdot \\ \qquad \cdot \left(f(x) - h \sum_{i=1}^{d} z^{(i)} \left.\frac{\partial f}{\partial z^{(i)}}\right|_x + \frac{h^2}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} z^{(i)} z^{(j)} \left.\frac{\partial^2 f}{\partial z^{(i)} \partial z^{(j)}}\right|_x + ...\right) dz = \\ \quad = h^{d+2} \sum_{i=1}^{d} \sum_{j=1}^{d} m_{2,i,j} A_{1,i,j}(x) + o\left(h^{d+2}\right) \end{array}$$

on account of condition H2 together with H5-H7.

Finally, it follows that:

$$\begin{aligned}
\alpha_3(x) &= \int\int K_d\left(\tfrac{x-u_1}{h}\right) K_d\left(\tfrac{x-u_2}{h}\right) \cdot \\
&\quad \cdot (C(u_2-u_1) - C(x-u_1) - C(x-u_2) + C(0)) \cdot \\
&\quad \cdot f(u_1)\, f(u_2)\; du_1 du_2 = \\
&= h^{2d} \int\int K_d(z_1)\, K_d(z_2)\, (C(hz_1 - hz_2) - C(hz_1) - C(hz_2) + C(0)) \cdot \\
&\quad \cdot f(x - hz_1)\, f(x - hz_2)\; dz_1 dz_2 = \\
&= h^{2d+1} \int\int K_d(z_1)\, K_d(z_2) \cdot \\
&\quad \cdot \left(2 \textstyle\sum_{i=1}^d z_2^{(i)} \left.\frac{\partial C}{\partial z^{(i)}}\right|_0 - h \sum_{i=1}^d \sum_{j=1}^d z_1^{(i)} z_2^{(j)} \left.\frac{\partial^2 C}{\partial z^{(i)} \partial z^{(j)}}\right|_0 + \ldots\right) \cdot \\
&\quad \cdot \left(f(x) - h \textstyle\sum_{i=1}^d z_1^{(i)} \left.\frac{\partial f}{\partial z^{(i)}}\right|_x + \ldots\right) \left(f(x) - h \textstyle\sum_{i=1}^d z_2^{(i)} \left.\frac{\partial f}{\partial z^{(i)}}\right|_x \ldots\right) dz_1 dz_2 = \\
&= -2h^{2d+2} \textstyle\sum_{i=1}^d \sum_{j=1}^d m_{1,i,j} A_{2,i,j}(x) + o\left(h^{2d+2}\right)
\end{aligned}$$

□

In view of Theorem 3.1, and having in mind the convergence rates established in hypothesis H7, one might conclude that the mean-squared prediction error of the proposed predictor is asymptotically zero.

Furthermore, the foregoing theorem allows us to derive an optimal bandwidth h, which would be selected so as to minimize the mean-squared prediction error above, which conveys that:

$$h_{opt}(x) = \left(\frac{(2-d)\sum_{i,j} m_{2,i,j} A_{1,i,j}(x)}{4\sum_{i,j} m_{1,i,j} A_{2,i,j}(x)}\right)^{1/d} n^{-1/d} \tag{2}$$

Remark 1 : The optimal h given in (2) may not be useful in practice. On one hand, h_{opt} turns out to be zero for $d = 2$. On the other, for $d \neq 2$, the optimal h is dependent on unknown quantities, whose estimation might be complex, such as the derivatives of the density of the random locations as well as those of the theoretical covariance function up to the first and second order, respectively.

Taking into account Remark **1**, two different approaches will be suggested for the selection of h which are intended to be easily applicable to real situations. The first one provides us with a global bandwidth, a fair compromise shared among all candidates to prediction locations all over the observation region. Its main advantage is that it can be achieved with small effort. Alternatively, we suggest some parametric bootstrap approach for the estimation of a local bandwidth, aiming to offer an optimal value depending on the location of each specific prediction point.

Our proposal to obtain an optimal global bandwidth is based on a classic cross-validation method. The fundamental idea behind this method, also called "the leave-one-out method", is to estimate $Z(x)$ at each sample point x_i from neighbouring data $Z(x_j),\ j \neq i,$ as if $Z(x_i)$ were unknown. In this way at every sample point x_i we get a prediction estimate, and the optimal value h_{glo} is the one which globally minimizes all prediction errors. So, given a data set $\mathbf{Z}$, the optimal global bandwidth is determined as follows:

$$h_{glo} = \text{argmin}_{h \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| Z(x_i) - \widehat{Z}^{-i}(x_i; h) \right| \right\} \tag{3}$$

where $\widehat{Z}^{-i}(x_i; h)$ represents the result of predictor NP at location x_i when removing $Z(x_i)$ and H is an adequate set of positive numbers when taking into account the

spatial distribution of the sample locations. Proceeding in this way, h_{glo} provides us with a bandwidth selector that may be applied for prediction at any given point x.

Our second proposal, designated for acquiring an optimal local bandwidth as a function of the prediction location x, will be based on a numerical Monte Carlo method. First, a satisfactory model for the sample data $\mathbf{Z}$ has to be chosen and, then, model parameters θ have to be estimated, for instance by some least squares criterium. Then, a large number of Monte Carlo simulated data sets should be generated, given $\hat{\theta}$, on the sample locations $x_1, ..., x_n$, as well as, on the new location x. Thus, suppose that we have a total of r replicas of $\mathbf{Z}$ denoted by $\mathbf{Z}^1, ..., \mathbf{Z}^r$. Then, the optimal local bandwidth can be computed as:

$$h_{loc}(x) = \text{argmin}_{h \in H} \left\{ \frac{1}{r} \sum_{i=1}^{r} \left| Z^i(x) - \widehat{Z}^i(x;h) \right| \right\} \tag{4}$$

where $\widehat{Z}^{\,i}(x;h)$ represents the result of predictor NP at location x when data $\mathbf{Z^i}$ are considered. As before, the optimal value $h_{loc}(x)$ given in (4) can then be used for predicting the value of Z at the given point x.

Remark 2 : In practice, H may be taken as a discrete set of positive equispaced values, depending on the minimum and maximum of the observed distances; the latter criterium will provide a global option for H. An alternative could be that of considering an upper bound for H given by the maximum distance from the selected location x, either to the boundary or to the spatial locations x_i. Moreover, H must be constructed so as to satisfy that a percentage of locations x_i are used in the implementation of $\widehat{Z}^i(x;h)$, for each x.

4. Simulation studies

We now describe some simulation studies done in order to analyse the performance of the prediction method suggested when adopting the optimal bandwidths given in (3) and (4). The results of these predictors were compared with those achieved with the ordinary kriging predictor (see e.g. [2]), under the assumptions of isotropy and second order stationarity, for both Gaussian and non-Gaussian data. In the latter case, we consider three distinct geostatistical processes: the chi-squared, the Poisson log-linear and the binomial logistic-linear.

A chi-squared process is easily obtained as $\{(S(x))^2 : x \in D\}$, where $S(x)$ is a Gaussian process with mean zero [2]. In this case, the covariance function of the new process will be given by $2C_S^2(u)$, where $C_S(u) = \sigma^2\rho(u)$ with σ^2 representing the variance of $S(x)$ and $\rho(.)$ its correlation function (Appendix A).

The other two geostatistical processes, which were considered in our simulation studies, correspond to the generalized linear geostatistical models presented in [4] and briefly described next.

4.1. *Generalized linear geostatistical models*

Suppose that our sample data are (x_i, Z_i), with $Z_i = Z(x_i)$ for $i = 1, ..., n$, and that our basic aim is to predict the realized values of an underlying spatial process $S(x)$. Moreover, note that $S(x)$ is a stationary Gaussian process with mean zero, variance σ^2 and correlation function ρ. If the random variables $Z_1, ..., Z_n$ are Gaussian, then

conditional on $S(x)$, they are mutually independent with conditional distributions:

$$[Z_i/S(x)] \sim N(\mu(x_i), \tau^2)$$

where $\mu(x) = \alpha + S(x)$, for some real-valued α, and τ^2 represents an unexplained non-spatial variation in Z, most likely some measurement error .

In our case, we wish to extend this conditional formulation of the Gaussian linear model to wider settings, more precisely we want to admit non-Gaussian probability distribution for each measurement Z_i conditional on $S(x)$. This class of models are usually referred to as generalized linear geostatistical models (GLGM). The GLGM generalize, and also include, the linear Gaussian model. The generic conditional expectations should now be represented as:

$$\mathrm{E}[Z_i/S(x)] = \mu(x_i) = g(\alpha + S(x_i))$$

where $g(.)$ is the analytic inverse of the link function $h(.)$ and α is some real value.

In our numerical studies, we consider the following GLGMs:

- The Gaussian model in which the link function simplifies to the identity function.

- The Poisson model in which the link function is the logarithm and the conditional ditribution of each Z_i is Poisson; so, one has

  $$[Z_i/S(x)] \sim Poisson(exp(\alpha + S(x_i))).$$

  This is a natural candidate model for spatially referenced count data. For example, suppose that $S(x)$ measures the surface of pollution in a given area, then Z_i could represent the concentration in ppm for a certain heavy metal. This example will be analysed in Section 4.2.

  The covariance function of the Poisson log-linear process may be approximated to $\exp(2\alpha)C_S(u)$, as checked in Appendix A.

- The binomial model in which the link function is the logit, and the measurements Z_i represent the outcomes of conditionally independent Bernoulli trials with $Prob[Z_i = 1/S(x)] = p(x_i)$. Therefore:

  $$\mathrm{E}[Z_i/S(x)] = \mu(x_i) = p(x_i) = \frac{exp(\alpha + S(x_i))}{1 + exp(\alpha + S(x_i))}.$$

  In practice, this model is more useful if the binary Z_i are replaced by conditionally binomial counts with large denominators n_i. An example is given in [3] where the aim was to analyse the spatial variation in the risk of a certain bacteria infection relative to other infections in a given region. The data was then treated as binomial counts at unit postcode locations, conditionally on an unobserved relative risk surface previously estimated.

  The binomial logistic-linear process has an approximate covariance function given by $\frac{\exp(2\alpha)}{(1+\exp(\alpha))^4}C_S(u)$, as we will show in Appendix A.

4.2. *Comparison of results*

In our studies, we took the symmetric Epanechnikov kernel, the observation region $D = [0, 4]^2 \subset \mathbb{R}^2$ and sample size $n = 100$. The data $Z_1, ..., Z_{100}$, Gaussian and

Table 1. Comparison of predictors P and NP through the means and standard deviations of the APE, derived from 100 simulated data sets of size 100.

Data	Method		MAPE	sd(APE)	$\overline{h}$	θ_S
Gaussian	P		1.0183	0.8776		$\mu_S = 8$
						$\sigma_S = 4$
		h_{glo}	1.2544	0.8208	0.69	$\phi_S = 2$
	NP	h_{loc}	**0.9982**	0.8102	0.48	$r = 10000$
Chi-squared	P		0.4502	0.4983		$\mu_S = 0$
						$\sigma_S = 1$
		h_{glo}	0.4847	0.5033	0.67	$\phi_S = 2$
	NP	h_{loc}	**0.4424**	0.4764	0.43	$r = 10000$
Poisson	P		3.0984	2.9543		$\mu_S = 2$
						$\sigma_S = 0.6$
		h_{glo}	3.0206	2.8825	0.97	$\phi_S = 2$
	NP	h_{loc}	**2.9807**	2.9416	0.87	$r = 50$
Binomial ($N = 20$)	P		2.4441	1.7983		$\mu_S = 0$
						$\sigma_S = 0.6$
		h_{glo}	2.3008	1.791	1.13	$\phi_S = 2$
	NP	h_{loc}	**2.2855**	1.7560	1.05	$r = 50$

non-Gaussian, were always generated from some underlying Gaussian process $S(x)$, with a covariance function given by an exponential model.

In a preliminary study, the impact of model misspecification of the covariance function was analysed. In this context, the NP predictor was compared with the parametric ordinary kriging predictor (from now on denoted as P), when considering for estimation two distinct covariance models: the matérn and the spherical ones. We have concluded that the effects of a wrong covariance model specification were present, but not that much significant. Consequently, the following studies did not consider the misspecification issue.

Table 1 summarizes the main results obtained on the comparison of P and NP predictors. In the latter case, results of a global and a local bandwidths, given in (3) and (4) respectively, are also compared. The absolute prediction error (APE) was adopted as a discrepancy measure between the true value and the predicted value. This simulation study has involved 100 simulated sample data sets, so the MAPE value identifies the mean of 10^4 APE, as each data set has sample size equals to 100.

For each distribution of the sample data, **bold** identifies the minimum MAPE value. In the case of the predictor NP, the mean value of the estimated bandwidths, denoted by $\bar{h}$, was also included in Table 1, aiming to represent the order of magnitude of the bandwidths, either local or global. Furthermore, note that the derivation of h_{loc} in (4) requires the simulation of new replicas for each dataset. In Table 1, r identifies the number of these new Monte Carlo simulated data sets. The parameters of $S(x)$, used to simulate Z_i, are given in the most right-handed column.

When analyzing Table 1, the main conclusion is that, for all four data types, the nonparametric method with a local bandwidth always produces the best results. Additionally, bear in mind that its main advantage, when compared to the P approach, is being less restrictive and not requiring restrictive theoretical assumptions which may queer the predictions. So, it should have a larger application field.

We can also observe that the chi-squared data presents the lowest means and standard deviations of APE, which seems to contradict the idea that these are typical Gaussian basic features. This is probably a side effect of a smaller σ_S being used for the the generation of this data.

According to our experience, the total number r of Monte Carlo simulated data sets, required by the nonparametric method in (4), should be as large as possible

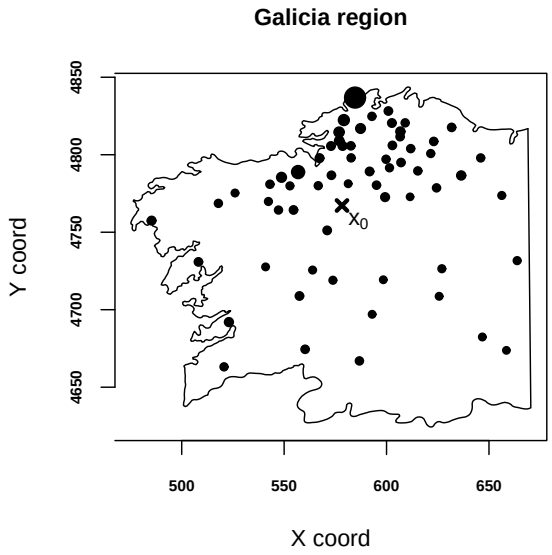

Figure 1. Spatial distribution of moss data in Galicia region.

depending on the computational cost one is prepared to pay. In the case of Gaussian and chi-squared data, the choice of this number is more demanding, so we have chosen $r = 10000$. This can be explained, taking in mind that it is very difficult to compete with the P method when data are Gaussian. The chi-squared process also shares this behaviour because its values are directly simulated from the Gaussian process. On the other hand, Poisson and binomial processes, which are less demanding, only depend on $S(x)$ process through its generation parameters λ and p, respectively. Moreover, note that we are applying the ordinary kriging directly on these data, with no data transformation, so it becomes easier to compete with the P method.

5. Application to real data

In this section, we present an application of the proposed predictors to a real data set. The data were collected from Galicia region in Spain for the analysis of air pollution intensities. Air quality can be monitored either by measuring the pollutants directly, providing objective but expensive information, or by using biomonitors, providing fast and inexpensive information. This last method is based on the high bioconcentration of heavy metals in land mosses. The typical procedure is to plant the moss and some time later to collect it to allow the concentration of heavy metals to be measured. More detail on this Galicia project of air pollution analysis can be found in [5].

The data set was collected in 1995 and it can be represented by (x_i, Z_i), where Z_i gives the concentrations of lead (Pb), measured in ppm, and collected in 63 locations x_i. Figure 1 illustrates the spatial distribution of moss data, noting that the bullets size is proportional to the corresponding Pb value. Some preliminary data analysis supports the stationarity assumption and, as expected with Poisson counts, the non-gaussianity of data.

First, we chose a random location x_0, also represented in Figure 1, as a new target of prediction. As in previous Section, the P and NP predictors, with a local and a

Table 2. Optimal bandwidths and predictions obtained at point x_0, for Galicia data set.

Method		$\hat{Z}_{x_0}$	h
P		3.1610	
	h_{glo}	4.6964	37.7
NP	h_{loc}	**3.0374**	22.2

Table 3. Means and standard deviations of APE and optimal bandwidths obtained for Galicia data set when applying a cross-validation method.

Method		MAPE	sd(APE)	$\overline{h}$
P		3.5941	7.1156	
	h_{glo}	3.2541	6.9214	37.7
NP	h_{loc}	**2.9715**	6.7782	22.1

global bandwidths, were used to obtain an estimate of the concentration of Pb on x_0. The results are summarized in Table 2. According to our simulation studies, the best estimate should be given by the nonparametric method when adopting a local bandwidth, which means that $Z(x_0) \approx 3.0374$ppm. Knowing that 50% of the locations have a concentration of Pb smaller than 3.69ppm, one can conclude that x_0 is not one of the locations with higher intensities of air pollution.

We have then proceeded with the assessment of the P and NP performance by applying the cross-validation method on $Z(x_i)$, $i = 1, ..., 63$. The cross-validation results are presented in Table 3, which confirm that the nonparametric method with an optimal local bandwidth continues to offer the best performance, with a h mean equals to 22.1km. So, also with this real data set, we have shown that nonparametric predictors can be preferable to the parametric predictor.

Appendix A. Covariograms of some non-Gaussian processes

This section presents the proofs of the relations between the covariogram of a Gaussian process $S(x)$ and the covariograms of non-Gaussian processes, generated from some transformation of $S(x)$.

A.1. *Chi-Squared processes*

If $\{S(x) : x \in D\}$ is a second-order stationary random process with mean μ, variance σ^2 and covariance function $C_S(.)$, then $\text{Cov}\left[(S(x_1))^2,\ (S(x_2))^2\right] = 2C_S^2(x_1 - x_2) + 4\mu^2 C_S(x_1 - x_2)$, for $x_1, x_2 \in D$.

Proof: From the properties of a normal distribution, one has:

(1) $\text{E}\left[S(x_1) - \mu\right] = 0$

(2) $\text{E}\left[\prod_{i=1}^{2}(S(x_i) - \mu)\right] = C_S(x_1 - x_2)$

(3) $\text{E}\left[\prod_{i=1}^{3}(S(x_i) - \mu)\right] = 0$

(4) $\text{E}\left[\prod_{i=1}^{4}(S(x_i) - \mu)\right] = C_S(x_1 - x_2)C_S(x_3 - x_4) + C_S(x_1 - x_3)C_S(x_2 - x_4) + C_S(x_1 - x_4)C_S(x_2 - x_3)$

where $x_1, x_2, x_3, x_4 \in D$.

In what follows, S_i is a shorthand notation for $S(x_i)$. Taking the stated

properties in mind and also that $\mathrm{E}\left[(S(x))^2\right] = \sigma^2 + \mu^2,\ \forall x \in D$, one has:

$$\begin{aligned}
\mathrm{Cov}\left[{S_1}^2,\ {S_2}^2\right] &= \mathrm{E}[\big({S_1}^2 - (\sigma^2+\mu^2)\big)\big({S_2}^2 - (\sigma^2+\mu^2)\big)] \\
&= \mathrm{E}[\big((S_1-\mu+\mu)^2 - (\sigma^2+\mu^2)\big)\big((S_2-\mu+\mu)^2 - (\sigma^2+\mu^2)\big)] \\
&= \mathrm{E}[\big((S_1-\mu)^2 + 2\mu(S_1-\mu) - \sigma^2\big) \\
&\qquad \times \big((S_2-\mu)^2 + 2\mu(S_2-\mu) - \sigma^2\big)] \\
&= \mathrm{E}[(S_1-\mu)^2(S_2-\mu)^2] - \sigma^4 + 4\mu^2\, C_S(x_1-x_2) \\
&= 2\, C_S^2(x_1-x_2) + 4\mu^2\, C_S(x_1-x_2)
\end{aligned}$$

□

In the particular case of $\mu = 0$, then $\mathrm{Cov}\left[(S(x_1))^2,\ (S(x_2))^2\right] = 2C_S^2(x_1-x_2)$ where $x_1, x_2 \in D$.

A.2. *Poisson and binomial processes*

Suppose that $\{S(x) : x \in D\}$ is a second-order stationary and isotropic Gaussian random process with zero mean, variance σ^2 and covariance function $C_S(.)$. Let observations $Z_i = Z(x_i),\ i = 1, \ldots, n$, conditional on $S(x)$, be mutually independent random variables with conditional expectations $\mu_i = g(\alpha + S_i)$ and conditional variances $v_i = v(\mu_i)$, where g is the analytic inverse of the link function L. So, for $u = \|x_i - x_j\| \in \mathbb{R}^+$ and for some real-valued α, the covariance function $C_Z(.)$ of the observed process $Z(x)$ can be approximated as follows.

(1) If $Z|S$ is a Poisson process, then $C_Z(u) \approx \exp(2\alpha) C_S(u)$
(2) If $Z|S$ is a binomial process, then $C_Z(u) \approx \frac{\exp(2\alpha)}{\left(1+\exp(\alpha)\right)^4} C_S(u)$

Proof: First, note that:

- $Z_i Z_j = \frac{1}{2}\left(Z_i^2 + Z_j^2 - (Z_i - Z_j)^2\right)$
- $\mathrm{E}[Z_i|S_i] = \mu_i = g(\alpha + S_i) \approx g(\alpha) + S_i g'(\alpha)$ (a first-order *Taylor* series approximation)
- $\mathrm{E_S}\left[v\big(g(\alpha+S_i)\big)\right] = \bar{\tau}^2$, meaning that the average of the conditional variance over the distribution of S is analogous to the nugget variance in the stationary Gaussian model.
- The following expression gives the relation between the variogram of the observed process $Z(.)$ and the variogram of the process $S(.)$.

$$\begin{aligned}
\gamma_Z(u) &= \tfrac{1}{2}\mathrm{E_Z}[(Z_i - Z_j)^2] = \tfrac{1}{2}\mathrm{E_S}\left[\mathrm{E_Z}[(Z_i - Z_j)^2|S]\right] \\
&= \tfrac{1}{2}\mathrm{E_S}\Big[E^2[(Z_i - Z_j)|S] + \mathrm{Var}[(Z_i - Z_j)|S]\Big] \\
&= \tfrac{1}{2}\mathrm{E_S}\left[\big(g(\alpha+S_i) - g(\alpha+S_j)\big)^2 + v\big(g(\alpha+S_i)\big) + v\big(g(\alpha+S_j)\big)\right] \\
&= \tfrac{1}{2}\mathrm{E_S}\left[\big(g(\alpha+S_i) - g(\alpha+S_j)\big)^2\right] + \mathrm{E_S}\left[v\big(g(\alpha+S_i)\big)\right] \\
&\approx \tfrac{1}{2}E\left[\big(g(\alpha) + S_i g'(\alpha) - g(\alpha) - S_j g'(\alpha)\big)^2\right] + \mathrm{E_S}[v_i] = [g'(\alpha)]^2\gamma_S(u) + \bar{\tau}^2
\end{aligned}$$

- $\mathrm{E_S}[\mathrm{E}[Z_i|S_i]] = \mathrm{E_S}[g(\alpha+S_i)] \approx \mathrm{E_S}[g(\alpha) + S_i g'(\alpha)] = g(\alpha) + g'(\alpha)\mathrm{E}[S_i] = g(\alpha)$
- $\mathrm{E_Z}[Z_i^2] = \mathrm{E_S}\left[\mathrm{E}[Z_i^2|S_i]\right] = \mathrm{Var_Z}[Z_i] + \left(\mathrm{E_S}[\mathrm{E}[Z_i|S_i]]\right)^2$

$$\begin{aligned}
&= \mathrm{E_S}[\mathrm{Var}[Z_i|S_i]] + \mathrm{Var_S}\left[\mathrm{E}[Z_i|S_i]\right] + \left[\mathrm{E_S}[\mathrm{E}[Z_i|S_i]]\right]^2 \\
&\approx \mathrm{E_S}[v_i] + \mathrm{Var_S}[g(\alpha) + S_i g'(\alpha)] + g^2(\alpha) \\
&= \bar{\tau}^2 + [g'(\alpha)]^2\mathrm{Var_S}[S_i] + g^2(\alpha) = \bar{\tau}^2 + [g'(\alpha)]^2\sigma^2 + g^2(\alpha)
\end{aligned}$$

Taking the above in mind, one has:

$$\begin{aligned} C_Z(u) &= C_Z(\|x_i - x_j\|) = \mathrm{Cov}\,[Z_i, Z_j] = \mathrm{E_Z}[Z_i Z_j] - \mathrm{E_Z}[Z_i]\,\mathrm{E_Z}[Z_j] \\ &= \tfrac{1}{2}\Big[\mathrm{E_Z}[Z_i^2] + \mathrm{E_Z}[Z_j^2] - \mathrm{E_Z}[(Z_i - Z_j)^2]\Big] - \mathrm{E_S}\big[\mathrm{E}[Z_i|S_i]\big]\ \mathrm{E_S}\big[\mathrm{E}[Z_j|S_j]\big] \\ &\approx 2 \times \tfrac{1}{2}\Big[\bar{\tau}^2 + [g'(\alpha)]^2\sigma^2 + g^2(\alpha)\Big] - [g'(\alpha)]^2\gamma_S(u) - \bar{\tau}^2 - g^2(\alpha) \\ &= [g'(\alpha)]^2(\sigma^2 - \gamma_S(u)) = [g'(\alpha)]^2 C_S(u) \end{aligned}$$

For the Poisson processes, $\log(\mu_i) = \alpha + S_i$. So, the link function L is the logarithm and the inverse of the link function is the exponential, i.e. $g(\alpha + S_i) = \exp(\alpha + S_i)$. Given this, one has $g'(\alpha) = \exp(\alpha)$. Thus, for the Poisson processes, the relation $C_Z(u) \approx \exp(2\alpha)\, C_S(u)$ immediately holds.

For the binomial processes, the link function L is the logit, and the responses $Z_i|S_i$ represent the outcomes of conditionally independent Bernoulli variables with expectations $\mu_i = p_i$, where $p_i = P[Z_i = 1|S_i]$. So, $L(\mu_i) = L(p_i) = \log(\frac{p_i}{1-p_i}) = \alpha + S_i$ and the inverse of the link function is $g(\alpha + S_i) = \mu_i = p_i = \frac{\exp(\alpha+S_i)}{1+\exp(\alpha+S_i)}$. Given this, one has $g'(\alpha) = \frac{\exp(\alpha)}{(1+\exp(\alpha))^2}$. Thus, for the binomial processes, the relation $C_Z(u) \approx \frac{\exp(2\alpha)}{(1+\exp(\alpha))^4}\, C_S(u)$ immediately holds.

□

References

[1] J. P. Chilés and P. Delfiner, *Geostatistics. Modeling spatial uncertainty*, Wiley Series in Probability and Statistics, Wiley, New York, 1999.
[2] N. Cressie, *Statistics for spatial data*, Wiley Series in Probability and Statistics, Wiley, New York, 1993.
[3] P.J. Diggle, J.A. Tawn and R.A. Moyeed, *Model-based Geostatistics*, Journal of the Royal Statistical Society, Series C, 47-3 (1998), pp. 299–350.
[4] P.J. Diggle and P.J. Ribeiro Jr., *Model-based Geostatistics*, Springer Series in Statistics, Springer, New York, 2007.
[5] J.A. Fernández, A. Rey and A. Carballeira, *An extended study of heavy metal deposition in Galicia (NW Spain) based on moss analysis*, Science of the Total Environment, 254 (2000), pp. 31–44.
[6] P. García-Soidán, M. Febrero and W. González-Manteiga, *Nonparametric kernel estimation of an isotropic semivariogram*, J. Statist. Plann. Inference, 121 (2004), pp. 65–92.
[7] L. Györfi, M. Kohler, A. Krzyzak and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*, Springer-Verlag, New York, 2002.
[8] P. Hall, I. Fisher and B. Hoffmann, *On the nonparametric estimation of covariance functions*, Ann. Statist., 22 (2004), pp. 2115–2134.
[9] P. Hall and J.S. Marron, *Estimation of integrated squared density estimators*, Statist. Probab. Letters, 6 (1987), pp. 109–115.
[10] P. Hall, J.S. Marron and B.U. Park, *Smoothed cross-validation*, Probab. Theory Rel. Fields, 92 (1992), pp. 1–20.
[11] W. Härdle, *Applied Nonparametric Regression*, Econometric Society Monographs, Cambridge University Press, Cambridge, 1990.
[12] B.U. Park and J.S. Marron, *On the use of pilot estimators in bandwidth selection*, J. Nonparametric Statist, 1 (1992), pp. 231–240.
[13] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistical Subjects, Chapman and Hall, London, 1986.
[14] M.P. Wand and M.C. Jones, *Kernel Smoothing*, Monographs on Statistics and Applied Probability, Chapman and Hall, London, 1995.