INSTITUT DE
STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

0832

LEAST SQUARES ESTIMATION OF NONLINEAR SPATIAL TRENDS

CRUJEIRAS, R. M. and I. VAN KEILEGOM

Least squares estimation of nonlinear spatial trends

Rosa M. Crujeiras * Ingrid Van Keilegom †

Abstract

The goal of this work is to study the asymptotic and finite sample properties of an estimator of a nonlinear regression function when errors are spatially correlated, and when the spatial dependence structure is unknown. The proposed method is based on a weighted nonlinear least squares approach, taking into account the spatial covariance. Weak consistency of the regression parameters estimator is derived, along with its asymptotic Gaussian limit. The behavior of the proposed estimator is illustrated with a simulation study, considering different correlation structures in $\mathbb{R}^2$ and a more general case including a spatial covariate. The method is also applied to two real data cases.

Key words and phrases. Asymptotic normality; Nonlinear least squares; Spatial regression; Variogram.

*Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium. E-mail address: rosa.crujeiras@usc.es. Research supported by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy) and Xunta de Galicia Project PGIDIT06PXIB207009PR.

†Institute of Statistics, Université catholique de Louvain, Voie du Roman Pays 20, B 1348 Louvain-la-Neuve, Belgium. E-mail address: ingrid.vankeilegom@uclouvain.be. Research supported by IAP research network grant nr. P6/03 of the Belgian government (Belgian Science Policy), and by the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013) / ERC Grant agreement No. 203650.

1 Introduction

Consider a regression model $Y(\mathbf{s}) = g(\mathbf{s}) + \epsilon(\mathbf{s})$, where $\mathbf{s}$ belongs to some subset of $\mathbb{R}^d$. The errors $\epsilon(\mathbf{s})$ are assumed to be spatially correlated, but we do not impose any type of structural condition. We are mostly interested in the case where $d = 2$ (spatial setting) or $d = 3$ (spatio-temporal setting), but other values of d are also possible. Spatial locations may be irregularly spaced. In this paper we suppose that the regression function g belongs to some parametric class $\{g_\theta : \theta \in \Theta\}$ of (nonlinear) functions, and we are interested in the estimation of the parameters of this model, when the error correlation structure is unknown.

Nonlinear trends are present in many examples involving spatially dependent data. For instance, in soil science, [21] describe a sigmoide growth curve for modelling the relationship between irrigation and soil water content; in environmental science, [15] considers a nonlinear trend model for sulphate deposition, and in meteorology, [24] derives a nonlinear model for acid deposition, based on a differential equation system.

In the context of nonlinear regression with time-dependent errors, [11] consider the estimation of a nonlinear regression function for time-dependent autoregressive errors. [25] obtain a more general result on the consistency and asymptotic normality of a nonlinear least squares estimator, with serially correlated time series errors.

For spatially correlated errors, [1] propose an estimation procedure for the parameters of a linear regression model, in the two-dimensional spatial case. The errors are assumed to follow a spatial unilateral first-order autoregressive moving average model and the regression parameters are obtained by an iterative procedure based on generalized least squares. The error covariance matrix is computed by maximum likelihood or a restricted maximum likelihood, based on residuals. However, no theoretical results are given.

In this work, we consider a more general case, where the trend may be nonlinear and the errors are spatially correlated, but without imposing any structural assumption. The method we propose in this work is a two-step procedure based on least squares estimation. The asymptotic distribution of the estimators is also obtained.

This paper is organized as follows. In the next section, we explain the precise estimation procedure for the regression parameters θ. The asymptotic properties of the proposed estimator are stated in Section 3, along with the conditions under which these properties are valid. In Section 4, an elaborated simulation study is carried out to study the finite sample performance of the estimators, whereas the estimation procedure is applied to two data sets in Section 5. Finally, the Appendix contains the proof of all asymptotic results.

2 The estimation procedure

When modelling a spatial process $Y(\mathbf{s})$, it is common to consider that

$$Y(\mathbf{s}) = g(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad D \subset \mathbb{R}^d \tag{2.1}$$

where g is the trend component, which captures the large-scale variability of the process, and $\epsilon(\mathbf{s})$ is a zero-mean second-order stationary process, describing the small-scale variability (see [6], Section 3.1). We consider a general parametric nonlinear model for the trend g, i.e.

$$g \in \mathcal{G} = \{g_\theta : \theta \in \Theta\},$$

with $\Theta \subset \mathbb{R}^p$ a compact set. Let θ_0 be the true value of θ, i.e. $g \equiv g_{\theta_0}$. The error process is assumed to be second-order stationary, so in particular intrinsic stationary (e.g. [6], p. 40). Hence we can describe the dependence structure via the variogram function, given by:

$$2\gamma(\mathbf{u}) = \mathrm{Var}(\epsilon(\mathbf{s}) - \epsilon(\mathbf{s}+\mathbf{u})), \quad \mathbf{s}, \mathbf{s}+\mathbf{u} \in D. \tag{2.2}$$

At fixed design points $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, we observe $Y(\mathbf{s}_i) = g(\mathbf{s}_i) + \epsilon(\mathbf{s}_i)$ $(i = 1, \ldots, n)$.

We are interested in the estimation of the parameter vector θ. The proposed estimation procedure consists of three steps: (1) unweighted least squares estimation of θ, ignoring the dependence structure of the errors; (2) estimation of the variance-covariance matrix of the errors based on the estimator of θ found in the first step; (3) weighted least squares estimation of θ, taking the dependence structure of the errors into account.

The proposed procedure is a generalization of the method proposed by [11] in the context of temporally autocorrelated errors. Our purpose is to adapt this method to a more general setting, for a spatial regression model with spatially dependent errors. The unknown dependence in the errors is not restricted to a spatial autoregression context (see [2]) and we do not impose a structural assumption (for instance, a Markov condition) on the spatial model. So, we go much further than simply extending the method from time-dependence to spatial dependence.

We now explain each of these three steps in detail. First, compute the ordinary least squares estimator for the regression parameter $\widetilde{\theta}$ as follows:

$$\widetilde{\theta} = \arg\min_{\theta \in \Theta} (\mathbf{Y} - \mathbf{g}_\theta)^t (\mathbf{Y} - \mathbf{g}_\theta),$$

where $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))^t$ and $\mathbf{g}_\theta = (g_\theta(\mathbf{s}_1), \ldots, g_\theta(\mathbf{s}_n))^t$.

In order to improve this preliminary estimator by taking into account the dependence structure of the errors, we will start with the estimation of the variogram. There is a

broad literature concerning the estimation of the variogram, either from parametric or nonparametric perspective. In this last context, variogram estimators are not generally valid, since they fail to satisfy the conditional negative definiteness property ([6], p. 86), or it is not easy to prove that this condition holds. However, nonparametric variogram estimators can be used as pilots for fitting a valid parametric model, by minimizing a certain criterion, such as the least squares distance.

The classical nonparametric estimator for the variogram is the empirical variogram, which is obtained by the method of moments:

$$2\widehat{\gamma}(\mathbf{u}) = \frac{1}{|N(\mathbf{u})|}\sum_{N(\mathbf{u})}(\widetilde{\epsilon}(\mathbf{s}_i) - \widetilde{\epsilon}(\mathbf{s}_j))^2, \tag{2.3}$$

where $N(\mathbf{u}) = \{(\mathbf{s}_i, \mathbf{s}_j) : \mathbf{s}_i - \mathbf{s}_j = \mathbf{u}\}$, $|N(\mathbf{u})|$ denotes the number of such pairs, and

$$\widetilde{\epsilon}(\mathbf{s}_i) = Y(\mathbf{s}_i) - g_{\widetilde{\theta}}(\mathbf{s}_i)$$

$(i = 1, \ldots, n)$ are the estimated errors. When the empirical variogram is computed from the (unobserved) errors $\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n)$ (instead of the residuals $\widetilde{\epsilon}(\mathbf{s}_1), \ldots, \widetilde{\epsilon}(\mathbf{s}_n)$), it is an asymptotically unbiased and consistent estimator of the variogram, in a pointwise sense (see [6], p.71), as long as the number of contributing pairs at each lag increases. However, this estimator may be influenced by outliers. For a Gaussian process, the squared differences involved in the latter empirical variogram follow a χ^2_1 distribution. Transforming these differences by a fourth-root, they are proved to have a distribution with skewness and kurtosis similar to the standard normal. Based on this argument, [4] proposed a robust version of this estimator.

When the observations are irregularly spaced, the variogram estimator (2.3) is usually smoothed by considering a tolerance region:

$$2\gamma^*(\mathbf{u}) = \frac{1}{|T(\mathbf{u})|}\sum_{T(\mathbf{u})}(\widetilde{\epsilon}(\mathbf{s}_i) - \widetilde{\epsilon}(\mathbf{s}_j))^2$$

where $T(\mathbf{u})$ is some specified neighbourhood of $\mathbf{u}$ in $\mathbb{R}^d$. Tolerance regions may be chosen small enough to keep the spatial dependence, but large enough to guarantee the stability of the estimator. For further discussion, see [6], p.70. In what follows, we will work with $2\widehat{\gamma}(\mathbf{u})$, but similar results can be obtained for the latter estimator $2\gamma^*(\mathbf{u})$.

Suppose that a valid parametric variogram family is given by $\{2\gamma_\alpha(\mathbf{u}) : \alpha \in A\}$, where A is a subset of $\mathbb{R}^q$. The parameter vector α can be estimated by using a weighted least squares approach, comparing the functions at lags $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ for some $K < \infty$:

$$\widehat{\alpha_{LS}} = \arg\min_{\alpha\in A} \widehat{\Gamma}(\alpha)^t W(\alpha)\widehat{\Gamma}(\alpha) := \arg\min_{\alpha\in A} \widehat{P}_n(\alpha), \tag{2.4}$$

where $\widehat{\Gamma}(\alpha) = (\gamma_\alpha(\mathbf{u}_1) - \widehat{\gamma}(\mathbf{u}_1), \ldots, \gamma_\alpha(\mathbf{u}_K) - \widehat{\gamma}(\mathbf{u}_K))^t$ and $W(\alpha)$ is an appropriate weight matrix. Ideally, we should take $W(\alpha) = \mathrm{Cov}(\widehat{\Gamma}(\alpha))^{-1}$, but it is not easy to find an expression for this covariance matrix, even for quite simple estimators of the variogram. When the errors would be observed (so θ would not need to be estimated in that case), and when the empirical estimator (2.3) or its robust version is used, the expression for the covariance matrix of $\widehat{\Gamma}(\alpha)$ has been derived by [5]. The strong consistency of the least squares estimators $\widehat{\alpha_{LS}}$ has been proved by [17] in that case, who also obtain the asymptotic distribution accounting for different sampling designs. For choosing the number of lags K, we may follow the recommendations from Journel and Huijbregths (see, for instance, [6], p. 70), with K fitting only up to half the maximum possible lag and considering only lags with $|N(\mathbf{u})|$ larger than 30, and $K \leq U/2$, where $U = \max\{\|\mathbf{u}\| : N(\mathbf{u}) > 0\}$.

Other fitting methods such as the maximum likelihood method could also be considered, in order to obtain a valid parametric variogram estimator. This estimation procedure relies on the Gaussian assumption, whereas the least squares method only depends on the asymptotic second-order structure of the process. Besides, the maximum likelihood parameter estimators may present a serious bias, although this problem can be mitigated using a restricted maximum likelihood approach.

Since the process $\epsilon(\mathbf{s})$ is second-order stationary, there exists a function C_α, called the covariogram, such that:

$$C_\alpha(\mathbf{u}) = \mathrm{Cov}(\epsilon(\mathbf{s}), \epsilon(\mathbf{s}+\mathbf{u})) = \sigma^2 - \gamma_\alpha(\mathbf{u}), \tag{2.5}$$

which can be recovered from the variogram, and where $\sigma^2 < \infty$ is the variance of the spatial process. For Gaussian processes, intrinsic and second-order stationarity are equivalent, although variogram estimation is usually preferred to covariogram estimation in practice. First, if the second-order stationarity of the process is not assessed, the covariogram C does not exists. In practice, this assumption can be checked using the test proposed by [10]. Another problem is that sample covariances do not provide unbiased estimators of the underlying covariances. This unbiasedness becomes a serious problem if there is any kind of trend 'contamination', which is the case here.

A valid estimator of the covariogram $C_\alpha(\mathbf{u})$ can be obtained by plugging-in in (2.5) the corresponding estimator of the variogram and a suitable estimator $\widehat{\sigma}^2$ of the variance. In most parametric variogram families, the variance parameter can be identified. In case this parameter can not be explicitly obtained from the model, we may use $(n-1)^{-1}\sum_{i=1}^n(\tilde{\epsilon}(\mathbf{s}_i) - \bar{\tilde{\epsilon}})^2$, as an estimator of the variance, where $\bar{\tilde{\epsilon}}$ denotes the average of the estimated errors. In the method proposed in this work, the variogram is computed based on residuals from a regression model. As it is pointed out in [16], the empirical variogram based on residuals is

seriously biased downwards, implying an underestimation of the variance when considering the method in (2.4). The authors monotonize the empirical variogram, applying the pool adjacent violators algorithm. The resulting empirical variogram is also strongly consistent.

Next, denote by $\Sigma = \Sigma_n$ the covariance matrix of the process $\{\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n)\}$, with entries:

$$\Sigma(i,j) = C_\alpha(\mathbf{s}_i - \mathbf{s}_j), \quad i,j = 1, \ldots, n.$$

This matrix can be estimated by $\widehat{\Sigma} = (\widehat{\Sigma}(i,j))_{i,j=1}^n$, where $\widehat{\Sigma}(i,j) = C_{\widehat{\alpha_{LS}}}(\mathbf{s}_i - \mathbf{s}_j)$. Since $\widehat{\Sigma}$ is an $n \times n$ symmetric and positive-definite matrix, the Cholesky decomposition allows to write:

$$\widehat{\Sigma} = \widehat{L}\widehat{L}^t,$$

where $\widehat{L}$ is a lower triangular $n \times n$ matrix.

Finally, define the following estimator of the regression parameter vector θ. This estimator (contrary to the preliminary estimator $\widetilde{\theta}$) is based on a weighted least squares criterion, which takes into account the dependence structure of the errors:

$$\widehat{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} (\widehat{L}^{-1}\mathbf{Y} - \widehat{L}^{-1}\mathbf{g}_\theta)^t (\widehat{L}^{-1}\mathbf{Y} - \widehat{L}^{-1}\mathbf{g}_\theta) := \arg\min_{\theta \in \Theta} Q_n(\theta). \tag{2.6}$$

Also note that although in the proposed method we consider a least squares procedure for the estimation of the variogram, it could be replaced by any other pointwise consistent estimator, as long as the result in Proposition 3.1 below is proved to hold.

An obvious question that arises is the construction of a second-step estimator for the dependence parameters, for instance, by a least squares criterion as in (2.4) based on $\widehat{\boldsymbol{\epsilon}} = Y - \mathbf{g}_{\widehat{\theta}}$. When estimating the variogram of a process based on residuals, we have to take into account the dependence structure in the data and the constraints required in the least squares procedure. Therefore, the two-stage estimator of α, obtained by replacing in the formula of $\widehat{\alpha_{LS}}$ the estimator $\widetilde{\theta}$ by $\widehat{\theta}$, will share with $\widehat{\alpha_{LS}}$ the same asymptotic properties, but the behavior on finite samples may not be satisfactory (see [6] or[12] for a detailed discussion). This problem has also been noticed by [11], for temporally autoregressive errors. In that context, improvements were only found when the autoregression condition was fullfilled and the correlation length was correctly identified.

3 Main results

Let $\nabla g_\theta = \left(\frac{\partial}{\partial\theta_1} g_\theta, \ldots, \frac{\partial}{\partial\theta_p} g_\theta\right)^t$ be the gradient of g_θ. Denote also by $G(\theta)$ the $n \times p$ Jacobian matrix of g_θ with respect to θ at the sampling points, i.e. the i-th row of $G(\theta)$ equals

$\nabla g_\theta(\mathbf{s}_i)^t$ $(i = 1, \ldots, n)$. The notation $\|A\| = \mathrm{tr}(A^t A)$ will be used for the Euclidean norm of any matrix A. We start by stating the assumptions under which the main asymptotic results will be valid.

Assumptions.

(A1) The spatial process $Y(\mathbf{s})$ can be represented as in (2.1), with $g_\theta \in \mathcal{G} = \{g_\theta : \theta \in \Theta\}$, $\Theta \subset \mathbb{R}^p$ a compact set. The error ϵ is a zero-mean, second-order stationary Gaussian process with covariance function $C_\alpha \in L^1(\mathbb{R}^d)$, $\alpha \in A$ with $A \subset \mathbb{R}^q$. The covariance function C_α is continuously differentiable with respect to α, $\sup\{|C_\alpha(\mathbf{u})| : \mathbf{u} \in D, \alpha \in A\} < \infty$ and $\sup\{|\frac{\partial}{\partial \alpha_j} C_\alpha(\mathbf{u})| : \mathbf{u} \in D, \alpha \in A, j = 1, \ldots, q\} < \infty$. Moreover, $|N(\mathbf{u}_i)|$ $(i = 1, \ldots, K)$ tends to infinity as n tends to infinity.

(A2) The weight matrix $W(\alpha)$ is positive definite and continuous for all $\alpha \in A$ and

$$\sup\{\|W(\alpha)\| + \|W(\alpha)\|^{-1} : \alpha \in A\} < \infty.$$

(A3) For all $\varepsilon > 0$, there exists a $\nu > 0$ such that $\inf_{\|\alpha - \alpha_0\| > \varepsilon} \sum_{i=1}^K (\gamma_\alpha(\mathbf{u}_i) - \gamma_{\alpha_0}(\mathbf{u}_i))^2 > \nu$.

(A4) The regression function $g_\theta(\mathbf{s})$ is differentiable with respect to the components of θ and

$$\max_{j=1,\ldots,p} \sup_{\theta \in \Theta, \mathbf{s} \in D} \left| \frac{\partial}{\partial \theta_j} g_\theta(\mathbf{s}) \right| < \infty.$$

(A5) For all $\varepsilon > 0$, there exists a $\nu > 0$ such that $\inf_{\|\theta - \theta_0\| > \varepsilon} R(\theta) > \nu$, where $R(\theta) = \lim_{n \to \infty} n^{-1} (\mathbf{g}_{\theta_0} - \mathbf{g}_\theta)^t \Sigma^{-1} (\mathbf{g}_{\theta_0} - \mathbf{g}_\theta)$, and $\mathbf{g}_\theta = (g_\theta(\mathbf{s}_1), \ldots, g_\theta(\mathbf{s}_n))$.

(A6) The sequence of matrices $\{S_n\}_n$, such that $S_n = n^{-1} G(\theta_0)^t \Sigma^{-1}(\alpha_0) G(\theta_0)$ converges to a positive definite matrix S as $n \to \infty$.

The Gaussianity assumption in $(A1)$ is required in order to apply the Law of Large Numbers and the Central Limit Theorem. This assumption could be relaxed by fixing the conditions under which the differences $\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)$, with $(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{u})$ are an associated sequence for a certain lag $\mathbf{u}$ (see [18]). Another alternative is to determine an appropriate Hájek-Rényi type inequality, and apply the results in [9]. In the proof of Propostion 3.2, the Gaussian assumption is a sufficient condition for applying a maximal inequality for degenerate U-processes (see [20]). Also in condition $(A1)$, considering an L^1-integrable covariance is not too restrictive, meaning that the error process $\epsilon(\mathbf{s})$ is short memory. The uniform bound condition on the weight matrix in $(A2)$ holds if the parameter space A is compact. Condition $(A3)$ is a form of identifiability condition on the parametric variogram model. This condition requires choosing the lag vectors $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$ in such a way that any two

variograms with different parameters take different values in at least one of the lags. That is to say, $2\gamma_{\alpha_1}$ and $2\gamma_{\alpha_2}$, with $\alpha_1 \neq \alpha_2$ can be distinguished by considering their values at $\{\mathbf{u}_1, \ldots, \mathbf{u}_K\}$. Similarly, condition $(A5)$ is required for identifying the θ-parameter.

We are now ready to state the main asymptotic results of the paper.

Proposition 3.1. *Assume that conditions* $(A1) - (A4)$ *hold. Then,*

$$\widehat{\alpha_{LS}} - \alpha_0 \to 0,$$

in probability, as $n \to \infty$.

Note that under conditions $(A1) - (A3)$, [17] proved that the estimator of α_0 obtained by replacing in the definition of $\widehat{\alpha_{LS}}$ the estimated errors $\widetilde{\epsilon}_i$ by the true errors ϵ_i $(i = 1, \ldots, n)$, converges a.s. to the true dependence parameter α_0. We are now ready to prove the weak consistency of $\widehat{\theta}$.

Proposition 3.2. *Assume that conditions* $(A1) - (A5)$ *hold. Then,*

$$\widehat{\theta} - \theta_0 \to 0,$$

in probability, as $n \to \infty$.

Theorem 3.3 below gives the asymptotic distribution of the estimator $\widehat{\theta}$ of the regression parameters.

Theorem 3.3. *Assume that conditions* $(A1) - (A6)$ *hold. Then, as* $n \to \infty$,

$$\sqrt{n}\,(\widehat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}_p(0, S^{-1}),$$

where S *is given in* $(A6)$.

Remark 3.4. Consider the following more general spatial regression model for the process $Y(\mathbf{s})$:

$$Y(\mathbf{s}) = g_\theta(X(\mathbf{s})) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in D, \quad D \subset \mathbb{R}^d, \tag{3.1}$$

where $X(\mathbf{s})$ is also a spatial varying process. This kind of models has become popular in practical applications, when combining two sources of information. For prediction purposes, model (3.1) motivates the so-called kriging method with external drift (see [3], Section 5.7.3), where the explanatory process $X(\mathbf{s})$ is usually observed at a finer grid, since it is necessary to observe it both at data locations and prediction locations. Just focusing on the estimation issue, which is the goal of this work, it can be shown that the above asymptotic results can be extended to this more general model, under suitable additional assumptions related to this new process $X(\mathbf{s})$, and taking into account the possible correlation between $X(\mathbf{s})$ and $Y(\mathbf{s})$.

4 Simulation study

In order to explore the performance of the estimation method proposed in Section 2, we have carried out a simulation study considering different scenarios for the spatial regression model (2.1). Data have been generated from an isotropic Gaussian spatial process $Y(\mathbf{s}) = Y(s_1, s_2)$ observed at regularly spaced locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in the unit square $[0,1]\times[0,1]$, with trend models:

$$g_1(\mathbf{s}) = \exp\left(\frac{s_1 s_2}{\theta}\right),$$
$$g_2(\mathbf{s}) = \theta \sin(2\pi s_1) + \cos(2\pi s_2).$$

We have also considered different isotropic dependence structures: exponential and spherical, with variance σ^2 and range parameter ϕ (that is $\alpha = (\sigma^2, \phi)^t$). For a second-order stationary process, the variance of the process is given by $\lim_{\|\mathbf{u}\|\to\infty} \gamma(\mathbf{u})$ (by relation (2.5)). The range is defined as the distance, in the isotropic context, beyond which observations become independent. The following semi-variograms γ are considered:

- Spherical: $\gamma_S(\mathbf{u}) = \sigma^2 \left(1.5\frac{\|\mathbf{u}\|}{\phi} - 0.5\left(\frac{\|\mathbf{u}\|}{\phi}\right)^3\right)^2$ for $\|\mathbf{u}\| \leq \phi$ and $\gamma_S(\mathbf{u}) = \sigma^2$, otherwise.
- Exponential: $\gamma_E(\mathbf{u}) = \sigma^2 \left(1 - \exp\left(-\frac{\|\mathbf{u}\|}{\phi}\right)\right)$.

In Table 1, we consider a spherical model for the variogram, and model g_1 for the trend. The mean, median and mean squared error (MSE) are calculated for the ordinary least squares estimator $\widetilde{\theta}$, obtained without taking into account the presence of a spatial dependence structure, and also for the proposed estimator $\widehat{\theta}$. Summary statistics for the estimators of the variance and the range parameter are also reported. It can be observed that in general the mean squared error of all estimators reduces significantly when the sample size is $n = 400$. For the estimation of the parameter θ, the second step estimator, $\widehat{\theta}$, clearly outperforms the first step estimator, $\widetilde{\theta}$. In Table 2 we consider the same trend g_1, but this time the variogram is exponential. The same conclusions can be drawn regarding the behaviour for increasing sample size, and regarding the comparison of $\widetilde{\theta}$ and $\widehat{\theta}$. Finally, Table 3 shows summary statistics for trend model g_2 and a spherical variogram. Again, the MSE of the estimator $\widehat{\theta}$ is smaller than that of the estimator $\widetilde{\theta}$, although the improvement found when increasing the sample size is not that large, maybe due to the complexity of the model.

For the more general regression model (3.1), we take $g_3(x) = \theta \sin(x)$, that is:

$$Y(\mathbf{s}) = g_3(X(\mathbf{s})) + \epsilon(\mathbf{s}),$$

	First step: $\theta = 0.5$		Second step: $\theta = 0.5$		$\sigma^2 = 1$		$\phi = 0.8$	
γ_S	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.512	0.515	0.507	0.500	1.053	0.949	0.865	0.821
Median	0.494	0.499	0.493	0.497	0.881	0.778	0.736	0.670
MSE	1.31e-04	2.89e-04	5.08e-05	1.42e-06	1.48e-01	4.20e-02	7.67e-02	4.83e-02
	First step: $\theta = 0.4$		Second step: $\theta = 0.4$		$\sigma^2 = 1$		$\phi = 0.5$	
γ_S	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.402	0.402	0.400	0.400	1.015	1.000	0.578	0.514
Median	0.398	0.403	0.399	0.397	0.955	0.978	0.497	0.441
MSE	3.24e-06	5.99e-06	7.05e-07	6.01e-08	1.22e-02	8.67e-03	1.84e-02	6.01e-03

Table 1: *Simulation results for trend model g_1 with spherical variogram. Mean, median and mean squared error (MSE) from 100 Monte Carlo experiments are reported for the first and second step estimators $\widetilde{\theta}$ and $\widehat{\theta}$, and for the estimators of σ^2 and ϕ.*

	First step		Second step	
	$\theta = 0.5$		$\theta = 0.5$	
γ_E, $(\sigma^2, \phi) = (1, 0.8)$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.562	0.534	0.504	0.498
Median	0.497	0.482	0.500	0.496
MSE	1.84e-02	2.04e-03	1.44e-05	2.96e-06
	$\theta = 0.25$		$\theta = 0.25$	
γ_E, $(\sigma^2, \phi) = (2, 0.5)$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.250	0.249	0.250	0.250
Median	0.250	0.249	0.250	0.249
MSE	1.47e-07	3.79e-07	1.10e-09	3.73e-10

Table 2: *Simulations results for trend model g_1 with exponential variogram. Mean, median and mean squared error (MSE) from 100 Monte Carlo experiments are reported for the first and second step estimators $\widetilde{\theta}$ and $\widehat{\theta}$.*

where $X(\mathbf{s})$ is a zero mean Gaussian spatial process with spherical variogram, variance $\sigma_X^2 = 1$ and range $\phi_X = 0.5$. In Table 4 we show the results for this model, considering two different parameter values. We observe the great improvement of considering the second step estimator $\widehat{\theta}$, in the reduction of the mean squared error, for different dependence models.

	First step		Second step	
	$\theta = 0.5$		$\theta = 0.5$	
γ_S, $(\sigma^2, \phi) = (1, 0.5)$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.520	0.544	0.510	0.507
Median	0.515	0.576	0.536	0.504
MSE	3.30e-02	8.41e-02	6.11e-03	4.56e-03
	$\theta = 0.25$		$\theta = 0.25$	
γ_S, $(\sigma^2, \phi) = (2, 0.8)$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.231	0.172	0.230	0.248
Median	0.277	0.246	0.245	0.300
MSE	0.201	0.213	3.20e-02	2.92e-02

Table 3: *Simulations results for trend model g_2 with spherical variogram. Mean, median and mean squared error (MSE) from 100 Monte Carlo experiments are reported for the first and second step estimators $\widetilde{\theta}$ and $\widehat{\theta}$.*

	First step		Second step	
	$\theta = 0.5$		$\theta = 0.5$	
γ_S, $(\sigma^2, \phi) = (1, 0.8)$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.526	0.518	0.487	0.499
Median	0.501	0.506	0.480	0.497
MSE	1.52e-02	1.63e-02	3.01e-04	1.05e-05
	$\theta = 0.25$		$\theta = 0.25$	
γ_S, $(\sigma^2, \phi) = (2, 0.8)$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
Mean	0.280	0.283	0.236	0.256
Median	0.234	0.294	0.218	0.255
MSE	9.31e-02	6.11e-02	8.07e-04	7.70e-05

Table 4: *Simulations results for trend model g_3, with external trend and spherical variogram. Mean, median and mean squared error (MSE) from 100 Monte Carlo experiments are reported for the first and second step estimators $\widetilde{\theta}$ and $\widehat{\theta}$. $X(\mathbf{s})$ is a zero mean Gaussian process with spherical variogram and $(\sigma^2_X, \phi_X) = (1, 0.5)$.*

5 Application to real data

In order to illustrate the performance of the method, we consider in this section two different data sets. The first data set collects surface elevations at different spatial locations, and is

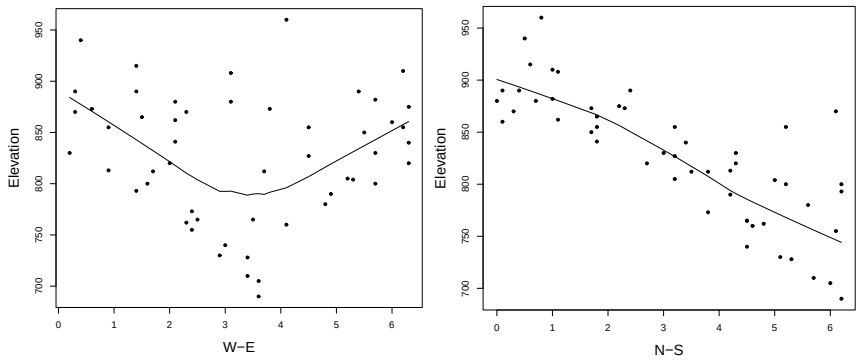

Figure 1: *Elevation against coordinates. Left plot: West-East direction. Right plot: North-South direction. Solid line: lowess curve.*

thoroughly described in [8], Chapter 1. The second data set is concerned with the Swiss Jura data from [14], which is available in the `R` package `gstat`. We apply the proposed algorithm to the first example, and we consider the extension for the more general model (3.1) in the second case.

5.1 Elevation data

Elevation models provide an ideal external drift function for mapping quantities which are controlled by topographic or orographic effects. For instance, digital elevation models have been used in order to model external trends for water level kriging (see [7]). Although our main goal is not to explore the issue of prediction with an external trend, it seems clear that a proper modelization of the random process with this collateral information will improve the kriging accuracy.

We have considered the elevation data studied in [8]. This dataset comprises a sample of elevation values at 52 locations within a square (sidelength of 6.7 units). The unit distances are 50 feet for the spatial locations, and 10 feet for elevation.

In Figure 1, we show the scatterplots of the elevation against the coordinates, in the West-East and North-South directions. This representation reveals the presence of a North-South trend, with higher values towards the eastern and western parts of the region. From this analysis, [8] consider two low-order polynomial trends: a linear and a quadratic trend model. We will check the performance of our method considering the linear trend model:

- Model 1: Linear trend $g_1(x, y) = \beta_0 + \beta_1 x + \beta_2 y$,

where x and y are, respectively, the coordinates in the West-East and North-South directions, and we compare our results with the ones obtained by [8]. From an ordinary least squares estimation (ignoring the spatial dependence structure), the estimated values for the parameters in Model 1 are: $\widehat{\beta}_0 = 913.80$, $\widehat{\beta}_1 = -1.695$ and $\widehat{\beta}_2 = -25.252$.

We have also considered a non-linear model, based on rational functions (see [19]) which have been proved to provide a more flexible approach than polynomials, with the same number of parameters. This type of rational functions has also been used by [15] for multivariate spatial prediction under non-stationarity and non-linear trend assumptions. The model considered is the following:

- Model 2: Rational trend $g_2(x, y) = \dfrac{1}{\theta_0 + \theta_1 x + \theta_2 y}$.

For the rational model, the estimated parameters by ordinary least squares are $\widehat{\theta}_0 = 1.08e-03$, $\widehat{\theta}_1 = 2.37e-06$ and $\widehat{\theta}_2 = 3.66e-05$.

For incorporating the spatial dependence, we have considered a Matérn covariance model. The Matérn covariance model is a three-parameter class of covariance functions, which in the isotropic case is given by

$$C_\alpha(u) = \sigma^2 \frac{(u/\phi)^\kappa K_\kappa\left(\frac{u}{\phi}\right)}{2^{\kappa-1}\Gamma(\kappa)}, \quad \alpha = (\kappa, \phi, \sigma^2),$$

where u is the distance between two sampling locations, σ^2 denotes the variance of the process, $K_\kappa(\cdot)$ is the modified Bessel function of the second kind of order κ, and $\phi > 0$ is a scale parameter which determines the rate at which the correlation decays to zero as u increases. The order parameter κ controls the analytic smoothness of the process. The value $\kappa = 0.5$ corresponds to the exponential covariance model, and $\kappa = 1.5$ and $\kappa = 2.5$ correspond to processes which are once and twice differentiable, respectively. Estimating the three parameters in the Matérn model is not a trivial task, since they are not well-identified. For that reason, [8] compares different values of a fixed smoothness parameter. We also follow this approach. For a more detailed description of the Matérn model and its properties, see [8], p.52 and [22], pp.32-33.

A histogram of the observed values shows a slight asymmetry and no presence of outliers, finding no reason for rejecting a Gaussian model. Based on this condition, [8] apply maximum likelihood estimation (MLE), assuming a Matérn covariance function with fixed smoothness parameter $\kappa = 1.5$. When a linear trend surface (Model 1) is considered, then the estimates are $\widehat{\beta}_0 = 912.5$, $\widehat{\beta}_1 = -16.5$ (West-East coordinate) and $\widehat{\beta}_2 = -5$ (North-South coordinate). The estimated values for the dependence parameters are, in this case,

	Model 1					
	$\widehat{\beta}_0$	$\widehat{\beta}_1$	$\widehat{\beta}_2$	$\widehat{\sigma}^2$	$\widehat{\phi}$	$\widehat{\tau}^2$
$\kappa = 0.5$	917.01	-17.73	-4.74	1557.57	1.45	0
$\kappa = 1.5$	949.62	-20.73	-3.55	3181.19	6.81	980.08
$\kappa = 2.5$	956.10	-21.92	-3.07	3121.69	4.70	1012.12

Table 5: *Parameter estimates with the proposed method, for the surface elevation data with linear trend model.*

	Model 3					
	$\widehat{\theta}_0$	$\widehat{\theta}_1$	$\widehat{\theta}_2$	$\widehat{\sigma}^2$	$\widehat{\phi}$	$\widehat{\tau}^2$
$\kappa = 0.5$	1.08e-03	7.08e-06	2.45e-05	1559.25	1.43	0
$\kappa = 1.5$	1.09e-03	6.73e-06	2.48e-05	1501.88	0.66	0
$\kappa = 2.5$	1.04e-03	4.76e-06	2.81e-05	3122.58	4.76	1023.14

Table 6: *Parameter estimates with the proposed method, for the surface elevation data with rational trend model.*

$\widehat{\sigma}^2 = 1693.1$, $\widehat{\tau}^2 = 34.9$ and $\widehat{\phi} = 0.8$, where τ^2 is called the nugget, and it appears as an additive term in the covariance expression (see [8], p.37 for discussion). The nugget effect has a dual interpretation, as the microscale variation of the process and as a measurement error variance. When the nugget is positive, then the variogram presents a discontinuity at the origin. The maximized log-likelihood is -240.08, for $\kappa = 1.5$.

The estimations obtained with the method proposed in this work, for the linear trend Model 1, are shown in Table 5, considering different smoothness parameters for the Matérn covariance. With the same number of parameters, we obtain the estimates for the rational trend model with normal errors (Model 2), which are shown in Table 6. For $\kappa = 1.5$, the maximized log-likelihood for the rational trend model is -242.8. For the linear trend model, we observe that the values of the estimated parameters are quite similar to those obtained by [8], using the MLE.

We have obtained kriging predictions for the surface elevation data, considering the rational trend. Results are shown in Figure 2. The prediction surface is quite similar to that provided by the constant and linear trend cases (see [8], p.38), but with different standard errors, running in this case between 0 and 27.61. The residuals for the linear trend model were reported between 0 and 22.9.

The estimates obtained for the regression parameters with the proposed method are quite

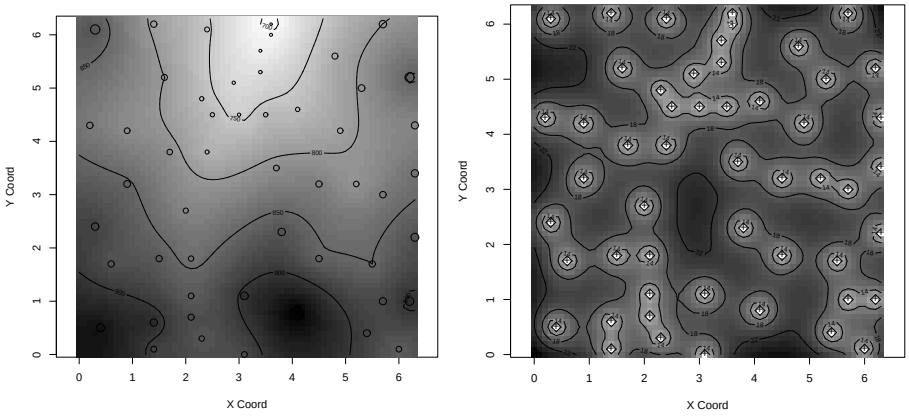

Figure 2: *Kriging with external drift. Left plot: kriging prediction, with rational trend model. Right plot: standard deviation predictions. The dots are the sampling locations.*

similar to those provided by the MLE. One of the advantages of the least squares approach compared with MLE is its computational efficiency. Another advantage is that, strictly speaking, the proposed method does not rely on the Gaussianity of the data, as long as the error differences form an associated sequence (see discussion after list of assumptions in Section 3).

5.2 Soil data

The Swiss Jura dataset contains measurements of heavy metal concentrations collected by the Swiss Federal Institute of Technology, in the Swiss canton of Jura. For a detailed description of the dataset, see [14]. For the illustration of the nonlinear trend estimation method in model (3.1), we consider 259 observations of cobalt (Co) and chromium (Cr), measured in mg/Kg.

Co is mainly derived from pollution sources, such as steelworks, whereas Cr may come both from pollution sources as well as from airborne particulates. In both cases, these metals present a close relation to indigenous soils. The relation between these two metals has already been reported by [26] and [13]. From an initial non-spatial descriptive analysis (see Figure 3), we observe a clear nonlinear relationship between Cr and Co.

We apply the method proposed in this paper to model the relation between these two metals, considering a sigmoidal curve:

$$Cr(\mathbf{s}) = \frac{\theta_0}{1 + \exp(\theta_1 - \theta_2 \cdot Co(\mathbf{s}))} + \epsilon(\mathbf{s}). \tag{5.1}$$

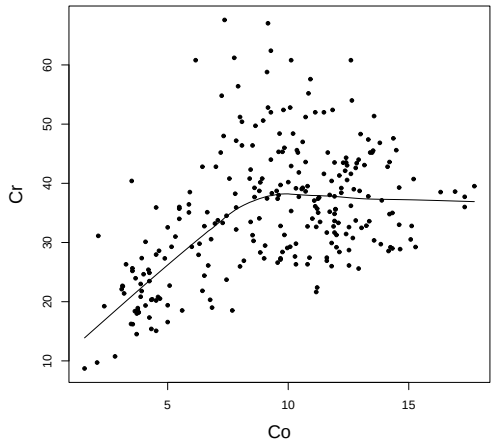

Figure 3: *Scatterplot of chromium against cobalt. Solid line: lowess curve.*

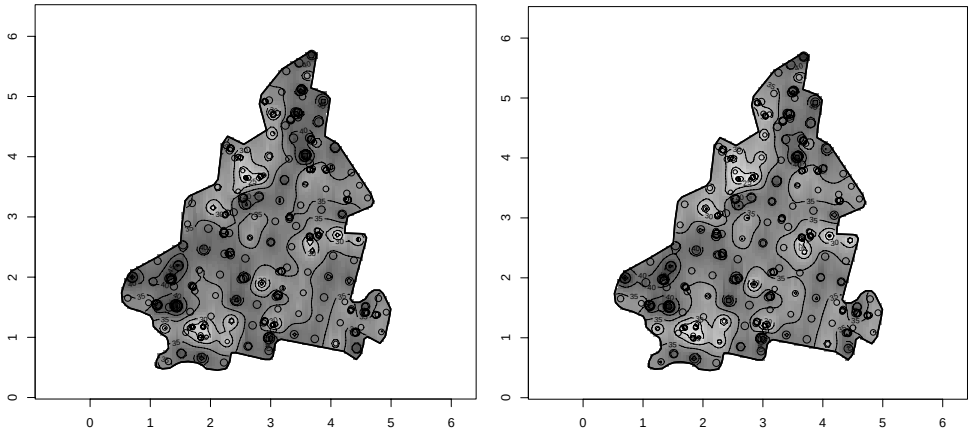

Figure 4: *Kriging surfaces for Cr. Left plot: ordinary kriging from original data. Right plot: ordinary kriging from estimated Cr values, using the proposed method with trend model (5.1). The dots are the sampling locations.*

The estimated parameters are $\widehat{\theta}_0 = 41.74$, $\widehat{\theta}_1 = 1.47$ and $\widehat{\theta}_2 = 0.40$. The first parameter represents the asymptote, that is, the maximum value of Cr achieved, for growing concentrations of Co. $\widehat{\theta}_1$ provides information about the intercept and $\widehat{\theta}_2$ is the rate at which the Cr concentration changes from its initial to its final value. We have assumed an exponential model for the dependence structure, with estimated parameters $\widehat{\phi} = 0.10$, $\widehat{\sigma}^2 = 62.92$ and nugget $\widehat{\tau}^2 = 13.11$.

We have also obtained in Figure 4 kriging surfaces for the Cr and for the estimated values of Cr, considering model (5.1). We can observe that the prediction surfaces are quite similar, indicating that the large scale variation of the Cr is well described by its nonlinear relationship with Co soil content.

6 Appendix

Proof of Proposition 3.1. For the sake of simplicity, we will restrict the proof to the ordinary least squares estimator. The extension to a weighted least squares problem is straightforward, taking into account assumption $(A2)$. Note that $\widehat{\alpha_{LS}}$ is the minimizer of the function $\widehat{P}_n(\alpha) = \sum_{i=1}^{K}(\gamma_\alpha(\mathbf{u}_i) - \widehat{\gamma}(\mathbf{u}_i))^2$. Define $P(\alpha) = \sum_{i=1}^{K}(\gamma_\alpha(\mathbf{u}_i) - \gamma_{\alpha_0}(\mathbf{u}_i))^2$. We prove the statement of the proposition by verifying the conditions of Theorem 5.7 in [23], p. 45) for $M_n = -\widehat{P}_n$ and $M = -P$, i.e. we will show that:

$$\text{For all } \varepsilon > 0, \exists \nu > 0 : \inf_{\|\alpha - \alpha_0\| > \varepsilon} |P(\alpha) - P(\alpha_0)| > \nu \tag{6.1}$$

$$\Delta_n = \sup_{\alpha \in A} |\widehat{P}_n(\alpha) - P(\alpha)| \xrightarrow{P} 0. \tag{6.2}$$

Condition (6.1) follows from assumption $(A3)$, since $P(\alpha_0) = 0$. By the proof of Theorem 3.1 in [17], we have that

$$\Delta_n^0 = \sup_{\alpha \in A} |P_n(\alpha) - P(\alpha)| \xrightarrow{P} 0,$$

where $P_n(\alpha) = \sum_{i=1}^{K}(\gamma_\alpha(\mathbf{u}_i) - \widetilde{\gamma}(\mathbf{u}_i))^2$ and $2\widetilde{\gamma}(\mathbf{u}) = |N(\mathbf{u})|^{-1} \sum_{N(\mathbf{u})}(\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j))^2$. We will prove that:

$$\sup_{\alpha \in A} |\widehat{P}_n(\alpha) - P_n(\alpha)| \xrightarrow{P} 0.$$

If we denote $\delta_\theta(\mathbf{s}) = g_{\theta_0}(\mathbf{s}) - g_\theta(\mathbf{s})$, we can write:

$$\begin{aligned}
2\widehat{\gamma}(\mathbf{u}) &= \frac{1}{|N(\mathbf{u})|}\sum_{N(\mathbf{u})}(\widetilde{\epsilon}(\mathbf{s}_i) - \widetilde{\epsilon}(\mathbf{s}_j))^2 = \frac{1}{|N(\mathbf{u})|}\sum_{N(\mathbf{u})}(\epsilon(\mathbf{s}_i) + \delta_{\widetilde{\theta}}(\mathbf{s}_i) - \epsilon(\mathbf{s}_j) - \delta_{\widetilde{\theta}}(\mathbf{s}_j))^2 \\
&= 2\widetilde{\gamma}(\mathbf{u}) + \frac{1}{|N(\mathbf{u})|}\sum_{N(\mathbf{u})}(\delta_{\widetilde{\theta}}(\mathbf{s}_i) - \delta_{\widetilde{\theta}}(\mathbf{s}_j))^2 + \frac{2}{|N(\mathbf{u})|}\sum_{N(\mathbf{u})}(\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j))(\delta_{\widetilde{\theta}}(\mathbf{s}_i) - \delta_{\widetilde{\theta}}(\mathbf{s}_j)) \\
&= 2\widetilde{\gamma}(\mathbf{u}) + 2A_n(\mathbf{u}) + 2B_n(\mathbf{u}),
\end{aligned}$$

where

$$\begin{aligned}
A_n(\mathbf{u}) &= \frac{1}{2|N(\mathbf{u})|}\sum_{N(\mathbf{u})}(\delta_{\widetilde{\theta}}(\mathbf{s}_i) - \delta_{\widetilde{\theta}}(\mathbf{s}_j))^2, \\
B_n(\mathbf{u}) &= \frac{1}{|N(\mathbf{u})|}\sum_{N(\mathbf{u})}(\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j))(\delta_{\widetilde{\theta}}(\mathbf{s}_i) - \delta_{\widetilde{\theta}}(\mathbf{s}_j)).
\end{aligned}$$

The function $\widehat{P}_n(\alpha)$ can now be decomposed as:

$$\begin{aligned}
\widehat{P}_n(\alpha) &= \sum_{i=1}^{K}(\gamma_\alpha(\mathbf{u}_i) - \widehat{\gamma}(\mathbf{u}_i))^2 = \sum_{i=1}^{K}(\gamma_\alpha(\mathbf{u}_i) - \widetilde{\gamma}(\mathbf{u}_i) - A_n(\mathbf{u}_i) - B_n(\mathbf{u}_i))^2 \\
&= P_n(\alpha) + \sum_{i=1}^{K}\left(A_n(\mathbf{u}_i) + B_n(\mathbf{u}_i)\right)^2 - 2\sum_{i=1}^{K}(\gamma_\alpha(\mathbf{u}_i) - \widetilde{\gamma}(\mathbf{u}_i))\left(A_n(\mathbf{u}_i) + B_n(\mathbf{u}_i)\right).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sup_\alpha |\widehat{P}_n(\alpha) - P_n(\alpha)| \\
&\leq \sum_{i=1}^{K}\left(A_n(\mathbf{u}_i) + B_n(\mathbf{u}_i)\right)^2 + \sup_\alpha \left|2\sum_{i=1}^{K}(\gamma_\alpha(\mathbf{u}_i) - \widetilde{\gamma}(\mathbf{u}_i))\left(A_n(\mathbf{u}_i) + B_n(\mathbf{u}_i)\right)\right|.
\end{aligned}$$

In order to prove that this quantity converges to zero in probability, it suffices to show that $A_n(\mathbf{u}_i) = o_P(1)$ and $B_n(\mathbf{u}_i) = o_P(1)$ for $i = 1, \ldots, K$. We will prove these conditions for any $\mathbf{u}$. We can write $A_n(\mathbf{u})$ as

$$\begin{aligned}
A_n(\mathbf{u}) &= \frac{1}{2|N(\mathbf{u})|}\sum_{N(\mathbf{u})}\left(\nabla G_{\widetilde{\theta}}^t(\mathbf{s}_i, \mathbf{s}_j)(\widetilde{\theta} - \theta_0)\right)^2 \\
&= (\widetilde{\theta} - \theta_0)^t \frac{1}{2|N(\mathbf{u})|}\sum_{N(\mathbf{u})}\nabla G_{\bar{\theta}}(\mathbf{s}_i, \mathbf{s}_j)\nabla G_{\widetilde{\theta}}^t(\mathbf{s}_i, \mathbf{s}_j)(\widetilde{\theta} - \theta_0),
\end{aligned}$$

where $\bar{\theta}$ is on the line segment between $\widetilde{\theta}$ and θ_0, and where

$$\nabla G_{\bar{\theta}}^t(\mathbf{s}_i, \mathbf{s}_j) = \left(\frac{\partial}{\partial\theta_1}g_\theta(\mathbf{s}_i) - \frac{\partial}{\partial\theta_1}g_\theta(\mathbf{s}_j), \ldots, \frac{\partial}{\partial\theta_p}g_\theta(\mathbf{s}_i) - \frac{\partial}{\partial\theta_p}g_\theta(\mathbf{s}_j)\right)\Bigg|_{\theta=\bar{\theta}}.$$

The convergence to zero in probability now follows from assumption $(A4)$ and from the fact that $\widetilde{\theta} \xrightarrow{P} \theta_0$ (see [11]). We can write $B_n(\mathbf{u})$ as:

$$\begin{aligned} B_n(\mathbf{u}) &= \frac{1}{|N(\mathbf{u})|} \sum_{N(\mathbf{u})} (\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)) \nabla G_{\bar{\theta}}^t(\mathbf{s}_i, \mathbf{s}_j)(\widetilde{\theta} - \theta_0) \\ &= (\widetilde{\theta} - \theta_0)^t \frac{1}{|N(\mathbf{u})|} \sum_{N(\mathbf{u})} (\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)) \nabla G_{\bar{\theta}}(\mathbf{s}_i, \mathbf{s}_j). \end{aligned}$$

For simplicity, we assume that $\dim(\Theta) = p = 1$. When $p > 1$, it suffices to prove this condition componentwise. Since $\widetilde{\theta}$ converges in probability to θ_0, we will prove that the remaining part of $B_n(\mathbf{u})$ is $O_P(1)$. Consider

$$\Big| \frac{1}{|N(\mathbf{u})|} \sum_{N(\mathbf{u})} (\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)) \nabla G_{\bar{\theta}}(\mathbf{s}_i, \mathbf{s}_j) \Big| \leq \frac{c}{|N(\mathbf{u})|} \sum_{N(\mathbf{u})} |\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)|,$$

where $c := 2 \sup_{\theta \in \Theta, \mathbf{s} \in D} |\frac{\partial}{\partial \theta} g_\theta(\mathbf{s})| < \infty$ by assumption $(A4)$. Also, note that by assumption $(A1)$, the absolute differences $|\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)|$, for $(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{u})$ follow a half-normal distribution with variance $2\gamma(\mathbf{u})(1 - 2/\pi)$. Hence, by Cauchy-Schwarz's inequality,

$$\begin{aligned} &\mathrm{Var}\left(\frac{1}{|N(\mathbf{u})|} \sum_{N(\mathbf{u})} |\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)| \right) \\ &\leq \frac{1}{|N(\mathbf{u})|^2} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{u})} \sum_{(\mathbf{s}_k, \mathbf{s}_l) \in N(\mathbf{u})} |\mathrm{Cov}(|\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)|, |\epsilon(\mathbf{s}_k) - \epsilon(\mathbf{s}_l)|)| \\ &\leq \frac{1}{|N(\mathbf{u})|^2} \sum_{(\mathbf{s}_i, \mathbf{s}_j) \in N(\mathbf{u})} \sum_{(\mathbf{s}_k, \mathbf{s}_l) \in N(\mathbf{u})} \sqrt{\mathrm{Var}(|\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)|) \mathrm{Var}(|\epsilon(\mathbf{s}_k) - \epsilon(\mathbf{s}_l)|)} \\ &\leq 2\gamma(\mathbf{u}) \left(1 - \frac{2}{\pi}\right) < \infty. \end{aligned}$$

It now easily follows that $|N(\mathbf{u})|^{-1} \sum_{N(\mathbf{u})} |\epsilon(\mathbf{s}_i) - \epsilon(\mathbf{s}_j)| = O_P(1)$, using Tchebychev's inequality. Hence, $B_n(\mathbf{u}) = o_P(1)$. □

Proof of Proposition 3.2. Similar to the proof of Proposition 3.1, we will check the following conditions:

$$\text{For all } \varepsilon > 0, \exists \nu > 0 : \inf_{\|\theta - \theta_0\| > \varepsilon} |Q(\theta) - Q(\theta_0)| > \nu \tag{6.3}$$

$$\Omega_n = \sup_{\theta \in \Theta} |Q_n(\theta) - Q(\theta)| \xrightarrow{P} 0, \tag{6.4}$$

where

$$Q(\theta) = 1 + R(\theta) = 1 + \lim_{n \to \infty} \frac{1}{n} \delta_\theta^t \Sigma^{-1} \delta_\theta,$$

and $\delta_\theta = (g_{\theta_0}(\mathbf{s}_1) - g_\theta(\mathbf{s}_1), \ldots, g_{\theta_0}(\mathbf{s}_n) - g_\theta(\mathbf{s}_n))^t$. Condition (6.3) follows from assumption $(A5)$ and the fact that $R(\theta_0) = 0$. For (6.4) note that the function $Q_n(\theta)$ can be decomposed as:

$$\begin{aligned} Q_n(\theta) &= \frac{1}{n}(\mathbf{Y} - \mathbf{g}_\theta)^t (\widehat{L}^{-1})^t \widehat{L}^{-1}(\mathbf{Y} - \mathbf{g}_\theta) \\ &= \frac{1}{n}(\epsilon + \delta_\theta)^t \widehat{\Sigma}^{-1}(\epsilon + \delta_\theta) = Q_{n1} + Q_{n2}(\theta) + Q_{n3}(\theta), \end{aligned}$$

where

$$Q_{n1} = \frac{1}{n}\epsilon^t \widehat{\Sigma}^{-1}\epsilon, \quad Q_{n2}(\theta) = \frac{2}{n}\epsilon^t \widehat{\Sigma}^{-1}\delta_\theta, \quad Q_{n3}(\theta) = \frac{1}{n}\delta_\theta^t \widehat{\Sigma}^{-1}\delta_\theta,$$

and $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$. Then,

$$\Omega_n \leq |Q_{n1} - 1| + \sup_{\theta \in \Theta} |Q_{n2}(\theta)| + \sup_{\theta \in \Theta} |Q_{n3}(\theta) - R(\theta)|.$$

For notational convenience, write $\widehat{\Sigma} = \Sigma(\widehat{\alpha_{LS}})$ and $\Sigma = \Sigma(\alpha_0)$. Note that

$$Q_{n1} = \frac{1}{n}\epsilon^t \Sigma^{-1}\epsilon + \sum_{j=1}^{q} (\widehat{\alpha_{LS}}_j - \alpha_{0j}) \frac{1}{n}\epsilon^t \frac{\partial \Sigma^{-1}(\widetilde{\alpha})}{\partial \alpha_j}\epsilon,$$

for some $\widetilde{\alpha}$ on the line segment between α_0 and $\widehat{\alpha_{LS}}$. To show that the second term on the right hand side is $o_P(1)$, first note that $\widehat{\alpha_{LS}}_j - \alpha_{0j} = o_P(1)$ for all $j = 1, \ldots, q$ by Proposition 3.1. Next, since $\widetilde{\alpha} - \alpha_0 = o_P(1)$, there exists a neighborhood $N(\alpha_0)$ of α_0 that contains $\widetilde{\alpha}$ with probability tending to one, and that is such that $\Sigma(\alpha)$ is positive definite for all $\alpha \in N(\alpha_0)$. We will prove that $\left(n^{-1}\epsilon^t \frac{\partial \Sigma^{-1}(\widetilde{\alpha})}{\partial \alpha_j}\epsilon\right)$ is $O_P(1)$ for $j = 1, \ldots, q$. For that purpose, note that ϵ is equal to Le, in distribution, for some $e^t = (e_1, \ldots, e_n) \sim N_n(0, I_n)$, where I_n is the $n \times n$ identity matrix. Then:

$$\begin{aligned} &\frac{1}{n}\left| e^t L^t \frac{\partial \Sigma^{-1}(\widetilde{\alpha})}{\partial \alpha_j} L e \right| \\ &\leq \frac{1}{n} \sup_{\alpha \in N(\alpha_0)} \left| e^t L^t \frac{\partial \Sigma^{-1}(\alpha)}{\partial \alpha_j} L e \right| \\ &\leq \frac{1}{n} \sup_{\alpha \in N(\alpha_0)} \left| \sum_{k=1}^{n} \left(L^t \frac{\partial \Sigma^{-1}(\alpha)}{\partial \alpha_j} L \right)_{kk} e_k^2 \right| + \frac{2}{n} \sup_{\alpha \in N(\alpha_0)} \left| \sum_{k<l} \left(L^t \frac{\partial \Sigma^{-1}(\alpha)}{\partial \alpha_j} L \right)_{kl} e_k e_l \right|. \end{aligned}$$

The first term on the right hand side is $O_P(1)$ by using similar arguments as for the term $B_n(\mathbf{u})$ in the proof of Proposition 3.1. For the second term, assumption $(A1)$ entails that the expression between absolute values is a degenerate U-process of order 2 indexed by a Euclidean class of functions with square integrable envelope function. Hence, it follows from Corollary 4 in [20] that the second term is also $O_P(1)$.

Now, consider

$$\frac{1}{n}\epsilon^t\Sigma^{-1}\epsilon = \frac{1}{n}(L^{-1}\epsilon)^t(L^{-1}\epsilon) \stackrel{d}{=} \frac{1}{n}\sum_{i=1}^{n} e_i^2,$$

where $e_1, \ldots, e_n$ are i.i.d. $N(0,1)$ random variables. Hence $n^{-1}\epsilon^t\Sigma^{-1}\epsilon \xrightarrow{P} 1$. This shows that $Q_{n1} - 1 = o_P(1)$. We now turn to $Q_{n2}(\theta)$. In a very similar way as for the term $B_n(\mathbf{u})$ in the proof of Proposition 3.1, we can show that $\sup_{\theta\in\Theta} |Q_{n2}(\theta)| = o_P(1)$. Finally, for $Q_{n3}(\theta)$, it is readily seen that $\sup_{\theta\in\Theta} |Q_{n3}(\theta) - R(\theta)| \xrightarrow{P} 0$, by using the weak consistency of $\widehat{\alpha_{LS}}$ established in Proposition 3.1. □

Proof of Theorem 3.3. Consider $Q_n(\theta) = Q_{n1} + Q_{n2}(\theta) + Q_{n3}(\theta)$, the same decomposition as in the proof of Proposition 3.2. Since $\widehat{\theta}$ is defined as the minimizer of $Q_n(\theta)$, we have that $\nabla Q_n(\widehat{\theta}) = 0$. If we examine the gradients of each addend in the decomposition of $Q_n(\theta)$, we see that $\nabla Q_{n1} = 0$ and

$$\nabla Q_{n2}(\theta) = \frac{2}{n}G(\theta)^t\Sigma^{-1}(\widehat{\alpha_{LS}})\epsilon, \quad \nabla Q_{n3}(\theta) = -\frac{2}{n}G(\theta)^t\Sigma^{-1}(\widehat{\alpha_{LS}})\delta_\theta,$$

where ϵ and δ_θ are defined as in the proof of Proposition 3.2. The vector $\nabla Q_{n3}(\theta)$ can be written after a first order Taylor expansion as

$$\nabla Q_{n3}(\theta) = -\frac{2}{n}G(\theta)^t\Sigma^{-1}(\widehat{\alpha_{LS}})G(\bar{\theta})(\theta - \theta_0),$$

for some $\bar{\theta}$ on the line segment between θ_0 and θ. Therefore, for $\theta = \widehat{\theta}$ we have the following equality :

$$G(\widehat{\theta})^t\Sigma^{-1}(\widehat{\alpha_{LS}})\epsilon = G(\widehat{\theta})^t\Sigma^{-1}(\widehat{\alpha_{LS}})G(\bar{\theta})(\widehat{\theta} - \theta_0).$$

It follows that

$$\sqrt{n}(\widehat{\theta} - \theta_0) = \sqrt{n}\left(G(\widehat{\theta})^t\Sigma^{-1}(\widehat{\alpha_{LS}})G(\bar{\theta})\right)^{-1} G(\widehat{\theta})^t\Sigma^{-1}(\widehat{\alpha_{LS}})\epsilon.$$

Now, note that by Propositions 3.1 and 3.2, the estimators $\widehat{\theta}$ and $\widehat{\alpha_{LS}}$ converge in probability to θ_0 and α_0. Hence, by Slutsky's theorem and noting that $\epsilon \stackrel{d}{=} Le$ with $e \sim N_n(0, I_n)$, it follows that the previous variable converges in distribution to a normal random variable with zero mean and covariance matrix S^{-1}, where

$$S = \lim_{n\to\infty} \frac{1}{n}G(\theta_0)^t\Sigma^{-1}(\alpha_0)G(\theta_0),$$

which is well-defined by assumption $(A6)$. □

References

[1] Basu, S. and Reinsel, G.C. (1994) Regression models with spatially correlated errors. *Journal of the American Statistical Association*, **89**, 88-99.

[2] Besag, J. (1974) Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, **36**, 192-236.

[3] Chiles, J.P. and Delfiner, P. (1999) *Geostatistics: Modeling Spatial Uncertainty.* Wiley, New York.

[4] Cressie, N. (1980) Robust estimation of the variogram. *Journal of the International Association for Mathematical Geology*, **12**, 115-125.

[5] Cressie, N. (1985) Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, **17**, 563-586.

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

[7] Desbarats, A.J., Logan, C.E., Hinton, M.J. and Sharpe, D.R. (2002) On the kriging of water table elevations using collateral information from a digitial elevation model. *Journal of Hydrology*, **255**, 25-38.

[8] Diggle, P. and Ribeiro, P.J. (2007) *Model-based Geostatistics.* Springer, New York.

[9] Fazekas, I. and Klesov, O. (2000) A general approach to the strong law of large numbers. *Theory of Probability and its Applications*, **45**, 569-583.

[10] Fuentes, M. (2005) A formal test for nonstationarity of spatial stochastic processes. *Journal of Multivariate Analysis*, **96**, 30-55.

[11] Gallant, A.R. and Goebel, J.J. (1976) Nonlinear regression with autocorrelated errors. *Journal of the American Statistical Association*, **71**, 961-967.

[12] Gambolati, G. and Galeati, G. (1987) Comments on Analysis of nonintrinsic spatial variability by residual kriging with application to regional groundwater levels. *Mathematical Geology*, **19**, 249-257.

[13] Gerdol, R., Bragazza, L. and Marchesini, R. (2002) Element concentrations in the forest moss *Hylocomium splendens*: variation associated with altitude, net primary production and soil chemistry. *Environmental Pollution*, **116**, 129-135.

[14] Goovaerts, P. (1997) *Geostatistics for Natural Resources Characterization.* Oxford University Press, New York.

[15] Haas, T. (1996) Multivariate spatial prediction in the presence of non-linear trend and covariance non-stationarity. *Environmetrics*, **7**, 145-165.

[16] Kim, H.J. and Boos, D.D. (2004) Variance estimation in spatial regression using non-parametric semivariogram based on residuals. *Scandinavian Journal of Statistics*, **31**, 387-401.

[17] Lahiri, S.N., Lee, Y. and Cressie, N. (2002) On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference*, **103**, 65-85.

[18] Matula, P. (1996) Convergence of weighted averages of associated random variables. *Probability and Mathematical Statistics*, **16**, 337-343.

[19] Ratkowsky, D.A. (1990) *Handbook of Nonlinear Regression Models*. Marcel Dekker, New York.

[20] Sherman, R.P. (1994) Maximal inequalities for degenerate U-processes with applications to optimization estimators. *Annals of Statistics*, **22**, 439-459.

[21] Snepvangers, J.J.J.C., Heuvelink, G.B.M. and Huisman, J.A. (2003) Soil water content interpolation using spatio-temporal kriging with external drift. *Geoderma*, **112**, 253-271.

[22] Stein, M.A. (1999) *Interpolation of Spatial Data*. Springer, New York.

[23] Van der Vaart, A.W. (1998) *Asymptotic Statistics*. Cambridge University Press, Cambridge.

[24] Verkatram, A. (1988) On the use of kriging in the spatial analysis of acid precipitation data. *Atmospheric Environment*, **22**, 1963-1975.

[25] White, H. and Domowitz, I. (1984) Nonlinear regression with dependent observations. *Econometrica*, **52**, 143-161.

[26] Zechmeister, H.G. (1995) Correlation between altitude and heavy metal deposition in the Alps. *Environmental Pollution*, **89**, 73-80.