Journal of Agricultural, Biological and Environmental Statistics manuscript No.
(will be inserted by the editor)

A distance-based method for spatial prediction in the presence of trend

Carlos E. Melo · Jorge Mateu · Oscar O. Melo

Abstract A new method based on distances for modeling continuous random data in Gaussian random fields is presented. In non-stationary cases in which a trend or drift is present, dealing with information in regionalized mixed variables (including categorical, discrete and continuous variables) is common in geosciences and environmental sciences. The proposed distance-based method is used in a geostatistical model to estimate the trend and the covariance structure, which are key features in interpolation and monitoring problems. This strategy takes full advantage of the information at hand due to the relationship between observations, by using a spectral decomposition of a selected distance and the corresponding principal coordinates. Unconditional simulations are performed to validate the efficiency of the proposed method under a variety of scenarios, and the results show a statistical gain when compared with a more traditional detrending method. Finally, our method is illustrated with two applications: earth's average daily temperatures in Croatia, and calcium concentration measured at a depth of 0-20 cm in Brazil.

Keywords Distance-based universal kriging · Gaussian random fields · Principal coordinates · Regionalized mixed variables · Unconditional simulations

Carlos E. Melo

Faculty of Engineering, Universidad Distrital Francisco José de Caldas

Tel.: +057-1-2053875

E-mail: cmelo@udistrital.edu.co

Jorge Mateu

Department of Mathematics, Universitat Jaume I, Campus Riu Sec, E-12071, Castellón, Spain

Tel.: +34-964-728391. Fax: +34-964-728429

E-mail: mateu@mat.uji.es

Oscar O. Melo

Department of Statistics, Faculty of Sciences, Universidad Nacional de Colombia

Tel.: +057-1-2053875

E-mail: oomelom@unal.edu.co

1 Introduction

Spatial information is collected and analyzed in a wide variety of scientific disciplines (mining, hydrogeology, ecology, earth sciences or environment). In the data analysis process, geostatistical interpolation methods such as simple, ordinary and universal kriging (UK) are widely known and used depending on the behavior of the trend associated with regionalized variables (see among others Cressie (1993), Jin et al. (2001), Joseph et al. (2008), Le & Zidek (2006), Sacks et al. (1989), Santner et al. (2003), van de Kassteele et al. (2009), Wackernagel (2003)). This paper is motivated by two real problems where we have a number of explanatory variables having an influence over a spatially dependent response variable. In the first application, the earth's average daily temperature in Croatia was measured at 154 meteorological stations (Hengl 2009), and the geographical locations (latitude and longitude), the Digital Elevation Model, the weighted topographic distance from the coastline and the Topographic Wetness Index were considered as explanatory variables. A second application concerns soil samples taken from a 0-20 cm depth layer at each of the 178 locations. Here the study region was divided into three sub-regions which experienced different soil management regimes, and calcium concentration and elevation were measured at a number of spatial locations (Capeche et al. 1997).

We are interested in those cases in which a trend or drift is present over the region of interest. The information coming from a set of explanatory variables has a direct influence and impact on the response variable. In practice, we often have to deal with explanatory variables of different nature that are associated with the response variable; these variables can be categorical or binary variables (e.g., type of soil or rock), or continuous variables (e.g., the spatial locations or additional environmental covariates). In the two applications presented here, the focus is on building a good model for spatial predictions, while it is not necessary to analyze the relationship that each explanatory variable has with the spatial response. In this way, we propose an alternative method for spatial interpolation based on distances that incorporates such valuable information when the explanatory variables are mixed (continuous, binary and categorical). Therefore, we use the Gower's distance (Gower 1968), although some other Euclidean distances may well be used. In our proposal, rather than the explanatory covariates entering the model directly as predictors, the principal coordinates are used, and can be applied in the geostatistical context to predict the trend and to estimate the the spatial correlation. Additionally, our proposed method produces better prediction performances than the classical universal kriging (UK).

Then, we develop a distance-based universal kriging (DBUK) model, which is an excellent alternative because it takes full advantage of the information obtained due to the relationship between observations. This can be established through the use of the spectral decomposition using any Euclidean distance. This approach aims at improving the predictions by allowing for inclusion of any number of principal

coordinates at the sample locations. Note that UK traditionally considers only information of a matrix of size $n \times p$, where n is the number of individuals and p the number of attributes with $p < n$. However, the proposed method considers a distance matrix between individuals with $n(n-1)/2$ different elements. Then, we build a matrix of principal coordinates using the spectral decomposition from this distance matrix getting information of $(n-1)$ independent columns, obtaining much more than p attributes for the trend. However, to avoid the problem of non-identifiability in the parameter estimation process for the trend in the DBUK proposed model, we select the most relevant principal coordinates (see Section 2), identifying which principal coordinates are most associated with the response variable. It is important to emphasize that principal coordinates transformation helps to reduce the noise in the geostatistical model because uncorrelated features are typically moved to higher order coordinates.

The method proposed models the mean part of a spatial model by covariates built from a distance measure matrix reduced by the principal coordinates related to the response variable. The associated UK prediction method is also developed allowing for prediction at unsampled locations, i. e. locations with missing measurements. The fact that our method is able to accommodate continuous and categorical variables is an attractive feature. Besides, compared with the usual regression trend, an important feature of our method is that it removes the need to test for interactions between predictors since the distance-based (DB) method used the spectral decomposition of the predictors and it captures the relationship between the predictors. This is an interesting fact since the definition of the trend model may have a major impact on the performance of a spatial model. Inference on model parameters and predictions are the usual aims of a spatial analysis. We thus propose a methodology to make predictions that involve mixed variables without considering a more classical non-parametric or semi-parametric method, in which including categorical and binary variables is more tricky. The proposed method is useful when the interest is not on the relationship between the explanatory variables and the spatial response. Thus the interpretability of the relation between covariates and response is somehow lost.

In the geostatistical regression case, the spatial DB method is based on methods developed by Cuadras (1989), Cuadras & Arenas (1990) and Cuadras et al. (1996). The DB is an excellent alternative to spatial interpolation with mixed explanatory variables because it takes full advantage of the information obtained due to the relationship between observations, which can be established through the use of the spectral decomposition using any Euclidean distance. This approach aims at improving the predictions by allowing for inclusion of any number of principal coordinates at the sample locations. Note that UK traditionally considers only information of a matrix of size $n \times p$, where n is the number of individuals and p the number of attributes with $p < n$; the DB proposed method considers a distance matrix between individuals with $n(n-1)/2$ different elements. Then, we build a matrix of principal coordinates using the

spectral decomposition from this distance matrix getting information of $(n-1)$ independent columns, obtaining much more than p attributes for the trend. To avoid the problem of non-identifiability in the parameter estimation process for the trend in the DB proposed model, we select the most relevant principal coordinates (see Section 2.1), identifying which principal coordinates are most associated with the response variable. It is important to emphasize that principal coordinates transformation helps to reduce the noise in the geostatistical model because uncorrelated features are typically moved to higher order coordinates.

Our spatial alternative based on DB methods for trend modeling is shown to be robust to misspecification in the correlation parameters (Cuadras et al. 1996). In this regard, unconditional simulations were performed to validate the efficiency of the proposed method under different conditions, and the results show a statistically significant gain when compared with the traditional UK model.

The plan of the paper is the following. Section 2 develops the methodological proposal, presents the variogram for the DB trend, shows the unbiasedness property of the predictor, and finally presents the UK built from the DB trend. Section 3 presents a simulation study based on Gaussian random field models. Section 4 develops the two applications that illustrate the proposed methodology. The paper ends with some conclusions.

2 A distance-based method

In this section, we obtain the principal coordinates from the matrix of similarities for a mixture of explanatory variables. Note that these coordinates can be also obtained from a matrix of Euclidean distances for continuous explanatory variables. Then, the principal coordinates that are the most correlated with the regionalized variable are selected using the criteria given in (7) and (8). Later, the variogram model using ordinary least squares (OLS), weighted least squares (WLS), maximum likelihood (ML) or restricted maximum likelihood (REML) is fitted, and the predictions at sample and unsampled points using DBUK are made. The prediction error variances are also calculated. Finally, the relation between the DBUK model and the classical geostatistical model is presented.

2.1 Distance-based trend

Suppose we are interested in a continuous geo-referenced response variable associated with binary, categorical or other continuous explanatory variables. Let $\boldsymbol{s} \in \mathbb{R}^d$ be a location in a d-dimensional Euclidean space, and suppose that $Z(\boldsymbol{s})$ is a random vector at each spatial location $\boldsymbol{s}$. If $\boldsymbol{s}$ is allowed to vary on a given set $D \subseteq \mathbb{R}^d$ (usually $d = 2$, but not necessarily), we have a stochastic process $\{Z(\boldsymbol{s}), \boldsymbol{s} \in D\}$,

which is object of study within the context of geostatistics (Cressie 1993). Assume that D is a certain fixed and continuous region, and the spatial index $\boldsymbol{s}$ varies continuously in D, i.e., there is an infinite number of possible locations where the process can be observed.

Assume the stochastic process follows a Gaussian random field model, which can be written as (Cressie 1993)

$$Z(\boldsymbol{s}_i) = \mu(\boldsymbol{s}_i) + \varepsilon(\boldsymbol{s}_i), \qquad i = 1, \ldots, n \tag{1}$$

where $Z(\boldsymbol{s}_i)$ is the regionalized variable given by the sum of a deterministic function associated with the trend $\mu(\boldsymbol{s}_i)$, and $\varepsilon(\boldsymbol{s}_i)$ is a zero-mean stationary stochastic component with variogram $2\gamma_\varepsilon(\cdot)$. Assume that the spatial trend is formed by categorical, continuous and binary variables, and it is modeled as

$$\mu(\boldsymbol{s}_i) = \theta_0 + \boldsymbol{v}'(\boldsymbol{s}_i)\boldsymbol{\theta}, \qquad i = 1, \ldots, n \tag{2}$$

where θ_0 is an unknown parameter, $\boldsymbol{v}(\boldsymbol{s}_i) = (v_1(\boldsymbol{s}_i), \ldots, v_p(\boldsymbol{s}_i))'$ is a vector containing explanatory variables associated to the spatial location $\boldsymbol{s}_i$, and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)'$ is a vector of unknown parameters.

In matrix form, model (1) can be expressed as

$$\boldsymbol{Z}_s = \mathbf{1}\theta_0 + \boldsymbol{V}\boldsymbol{\theta} + \boldsymbol{\varepsilon}_s \tag{3}$$

where $\boldsymbol{Z}_s = (Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))'$, $\mathbf{1}$ is a vector of dimension $n \times 1$ associated with the intercept, $\boldsymbol{V} = (\boldsymbol{V}_1, \ldots, \boldsymbol{V}_p)$ is a design matrix of dimension $n \times p$ with p explanatory variables given by $\boldsymbol{V}_j = (v_j(\boldsymbol{s}_1), \ldots, v_j(\boldsymbol{s}_n))'$, $j = 1, \ldots, p$, and $\boldsymbol{\varepsilon}_s = (\varepsilon(\boldsymbol{s}_1), \ldots, \varepsilon(\boldsymbol{s}_n))'$. By simplicity, we assume that $\boldsymbol{V} = \boldsymbol{H}\boldsymbol{V}^*$ where $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^t$ is the centering matrix with $\boldsymbol{I}$ the identity matrix of size $n \times n$. $\boldsymbol{V}^*$ is the matrix of original explanatory variables, it may involve continuous, categorical and binary variables, or even a mixture of them. Note that it is not necessary to centralize the covariates in the proposed model; this is performed with the purpose to build the distances, but depending on the distance used we do not always need to centralize the covariates.

Our distance-based approach applies to a transformation of the explanatory variables. To this aim, we need to define some similarity (or Euclidean distance) measures, which depend on the explanatory variable characteristics:

1. If the vector $\boldsymbol{v}(\boldsymbol{s}_i)$ given in (2) is formed by binary, categorical and continuous variables, the similarity according to Gower (1971) can be defined for mixed variables as

$$m_{ii'} = \frac{\sum_{h=1}^{p_1}\left(1 - \frac{|v_h(\boldsymbol{s}_i) - v_h(\boldsymbol{s}_{i'})|}{G_h}\right) + c_{1ii'} + \omega_{ii'}}{p_1 + (p_2 - c_{4ii'}) + p_3} \qquad i, i' = 1, \ldots, n \tag{4}$$

where G_h is the range of the h-th continuous variable, p_1 is the number of continuous variables, $c_{1ii'} = c_1(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$ and $c_{4ii'} = c_4(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$ are the number of positive and negative matches, respectively,

for p_2 binary variables associated with the relationship between $\boldsymbol{s}_i$-th and $\boldsymbol{s}_{i'}$-th locations, and $\omega_{ii'} = \omega(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$ is the number of matches for p_3 multistate variables. Through the transformation

$$\delta^2_{ii'} = m_{ii} + m_{i'i'} - 2m_{ii'} = 2(1 - m_{ii'})$$

it is possible to obtain Euclidean distances.

2. If all explanatory variables are binary or categorical, the similarity reduces to (Sokal & Michener 1958)

$$m_{ii'} = \frac{c_{1ii'} + c_{4ii'}}{c_{1ii'} + c_{2ii'} + c_{3ii'} + c_{4ii'}}$$

which is the number of agreements (matches) divided by the total number in the crossed table of i and i', and where $c_{1ii'}$, $c_{2ii'} = c_2(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$, $c_{3ii'} = c_3(\boldsymbol{s}_i, \boldsymbol{s}_{i'})$, $c_{4ii'}$ are the frequencies of (1,1), (1,0), (0,1) and (0,0), respectively. Note that for l categories of one categorical variable, $l-1$ dummy variables are created of ones and zeros corresponding to the presence or absence of one category (e.g. with 1 for the presence of A category in that variable and zero for the non-presence of that category, and we did the same procedure for $B, C, \ldots$ categories). Therefore, the Sokal - Michener equation can be applied because we have $l-1$ binary variables. The conversion $\delta^2_{ii'} = 2(1 - m_{ii'})$ yields Euclidean distances in this case.

3. When explanatory variables are continuous in (2), we can use the Euclidean distance given by

$$\begin{aligned}\delta^2_{ii'} =& (\boldsymbol{v}(\boldsymbol{s}_i) - \boldsymbol{v}(\boldsymbol{s}_{i'}))'(\boldsymbol{v}(\boldsymbol{s}_i) - \boldsymbol{v}(\boldsymbol{s}_{i'})) \\ =& \boldsymbol{v}'(\boldsymbol{s}_i)\boldsymbol{v}(\boldsymbol{s}_i) + \boldsymbol{v}'(\boldsymbol{s}_{i'})\boldsymbol{v}(\boldsymbol{s}_{i'}) - 2\boldsymbol{v}'(\boldsymbol{s}_i)\boldsymbol{v}(\boldsymbol{s}_{i'}) \end{aligned} \tag{5}$$

Expressions for the Gower similarity as given in equation (4) will be useful to the extent of having information associated with mixed variables, not only for the sampling sites but also for the non-sampled, which limits its use in unsampled areas.

After selecting one of the previously commented distances, let $\boldsymbol{A}_{n\times n} = (a_{ii'})$ be the matrix with elements $a_{ii'} = -\delta^2_{ii'}/2$ and $\boldsymbol{B} = \boldsymbol{HAH}$ where $\boldsymbol{H} = \boldsymbol{I} - \frac{1}{n}\mathbf{1}\mathbf{1}'$ is the centering matrix with $\boldsymbol{I}$ the identity matrix of size $n \times n$, and $\mathbf{1}$ a $n \times 1$ vector of ones. It is known that $\boldsymbol{B}$ is a semi-definite positive matrix (Mardia et al. 2002) of rank $n-1$. The matrix $\boldsymbol{X}$ of principal coordinates can be obtained from the spectral decomposition as

$$\boldsymbol{B} = \boldsymbol{HAH} = \boldsymbol{U\Lambda U}' = \boldsymbol{XX}'$$

where $\boldsymbol{\Lambda}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{B}$, $\boldsymbol{X} = \boldsymbol{U\Lambda}^{1/2}$ is an $n \times n$ matrix of rank $n-1$ because it has an eigenvalue equal to $\mathbf{1}$, and $\boldsymbol{U}$ contains the standardized coordinates.

Thus, given only the most significant principal coordinates, we can write the distance-based model as

$$\boldsymbol{Z}_s = \mathbf{1}\beta_0 + \boldsymbol{X}_{(k)}\boldsymbol{\beta} + \boldsymbol{e}_s \tag{6}$$

where $\boldsymbol{X}_{(k)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ contains a subset of $k \leq (n-1)$ relevant columns (principal coordinates) of $\boldsymbol{X}$ that are significantly correlated with $\boldsymbol{Z}_s$, β_0 and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ are the unknown parameters, and $\boldsymbol{e}_s = (e(\boldsymbol{s}_1), \ldots, e(\boldsymbol{s}_n))'$, with each $e(\boldsymbol{s}_i)$ a spatial error which is a zero mean intrinsically stationary process with variogram $2\gamma_\varepsilon(\cdot)$. Moreover, note that $\mathbf{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k$ are eigenvectors of $\boldsymbol{B}$ with eigenvalues $0, \lambda_1, \ldots, \lambda_k$, respectively, and $\boldsymbol{X}_j'\boldsymbol{X}_j = \lambda_j$, $\boldsymbol{X}_j'\boldsymbol{X}_{j'} = 0$ $(j \neq j')$, and $\boldsymbol{X}_j'\mathbf{1} = 0$, $j, j' = 1, \ldots, k$.

To avoid the problem of having a coefficient of determination $R^2 \simeq 1$ when the rank of $\boldsymbol{X}$ is large (k close to $n-1$), it is necessary to consider only those eigenvectors of $\boldsymbol{B}$, given by $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$, that are the most significatly correlated with the regionalized variable $\boldsymbol{Z}_s$, i.e., the most significantly correlated principal coordinates with $\boldsymbol{Z}_s$. In order to decide whether a predictor variable, i.e., a column of $\boldsymbol{X}$, has to be included or deleted, the variables can be arranged in decreasing order of the square of the correlation coefficients with $\boldsymbol{Z}_s$,

$$r^2(\boldsymbol{Z}_s, \boldsymbol{X}_1) > \cdots > r^2(\boldsymbol{Z}_s, \boldsymbol{X}_k) > r^2(\boldsymbol{Z}_s, \boldsymbol{X}_{k+1}) > \cdots > r^2(\boldsymbol{Z}_s, \boldsymbol{X}_{n-1}) \tag{7}$$

where $r^2(\boldsymbol{Z}_s, \boldsymbol{X}_j) = \dfrac{\boldsymbol{Z}_s'\boldsymbol{X}_j\boldsymbol{X}_j'\boldsymbol{Z}_s}{\lambda_j \sum_{j'=1}^{n} \left[Z(\boldsymbol{s}_{j'}) - \bar{Z}\right]^2}$, $j = 1, \ldots, n-1$, with $\bar{Z} = \sum_{j'=1}^{n} Z(\boldsymbol{s}_{j'})/n$.

Moreover, a principal coordinate $\boldsymbol{X}_j$ should be deleted if the null hypothesis $\beta_j = 0$ is not rejected. Following Cuadras & Arenas (1990), we can build a statistical test of the form

$$t_j = \frac{\hat{\beta}_j}{\|\boldsymbol{Z}_s - \hat{\beta}_0\mathbf{1} - \boldsymbol{X}\hat{\boldsymbol{\beta}}^*\|^2}\sqrt{\lambda_j(n-k-1)}, \qquad j = 1, \ldots, n-1 \tag{8}$$

where $\hat{\beta}_0 = \bar{Z}$, $\hat{\boldsymbol{\beta}}^* = \boldsymbol{\Lambda}^{-1}\boldsymbol{X}'\boldsymbol{Z}_s$ and $\hat{\beta}_j$ is the j-th component of $\hat{\boldsymbol{\beta}}^*$. To ease statistical inference, it is commonly assumed that the zero-mean residual part is multivariate normally distributed, so that we can approximate (8) to a t-student distribution with $(n-k-1)$ degrees of freedom.

Other possibilities are the largest eigenvalues (see details in Cuadras (1993) and Cuadras et al. (1996)) and the predictability criterion (see details in Cuadras et al. (1996)). Finally, following any of the methods presented above, the $\boldsymbol{X}_{k+1}, \ldots, \boldsymbol{X}_{n-1}$ principal coordinates must be removed.

2.2 Distance-based universal kriging

In Section 2.1, we have described how principal coordinates can be obtained from the spectral decomposition of the matrix $\boldsymbol{B}$, which is obtained as a transformation of the similarity (or distance) matrix calculated from explanatory variables. Then, these principal coordinates are used to estimate the trend component of equation (6). Note that UK traditionally considers only information of a matrix of p attributes with $p < n$, while the DB method considers k independent principal coordinates with $p \leq k \leq n-1$, thus obtaining more attributes in our DB method. If $k \leq p$, the DB method should not be

used because there is a loss of information. However, if this loss is small, the users of this methodology will obtain similar results to the traditional UK.

Once the trend has been estimated using the DB method, we can detrend the data and obtain the residuals $\boldsymbol{e}_s$ in (6). We now first describe the fitting of the variogram of the residuals from the trend estimated using the DB method. The experimental semivariogram $\widehat{\gamma}(h)$ is a key tool to any geostatistical analysis because it describes the spatial correlation of the regionalized variable at different distances. A natural estimator based on the method of moments, due to Matheron (Cressie 1993), is given by

$$2\hat{\gamma}(\boldsymbol{h}) = \frac{1}{|N(\boldsymbol{h})|} \sum_{N(\boldsymbol{h})} [\hat{e}(\boldsymbol{s}_i) - \hat{e}(\boldsymbol{s}_{i'})]^2$$

where $N(\boldsymbol{h}) \equiv \{(i, i') : \boldsymbol{s}_i - \boldsymbol{s}_{i'} = \boldsymbol{h}\}$ and $|N(\boldsymbol{h})|$ is the number of distinct elements of $N(\boldsymbol{h})$, $\hat{e}(\boldsymbol{s}_i) = Z(\boldsymbol{s}_i) - \hat{Z}(\boldsymbol{s}_i)$ and $\hat{e}(\boldsymbol{s}_{i'}) = Z(\boldsymbol{s}_{i'}) - \hat{Z}(\boldsymbol{s}_{i'})$ are the residual values at locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$, respectively, with $\hat{Z}(\boldsymbol{s}_i)$ and $\hat{Z}(\boldsymbol{s}_{i'})$ the predictions at locations $\boldsymbol{s}_i$ and $\boldsymbol{s}_{i'}$ using the DB method. This estimator is generally biased in presence of atypical information, affecting the estimate. Cressie & Hawkins (1980) proposed a robust estimator given by

$$2\bar{\gamma}(\boldsymbol{h}) = \frac{1}{0.457 + 0.494/|N(\boldsymbol{h})|} \left(\frac{1}{|N(\boldsymbol{h})|} \sum_{(\boldsymbol{s}_i, \boldsymbol{s}_{i'}) \in N(\boldsymbol{h})} |\hat{e}(\boldsymbol{s}_i) - \hat{e}(\boldsymbol{s}_{i'})|^{1/2} \right)^4 \tag{9}$$

They suggested a fourth-root transformation which led to the constants given in the above Equation (9) (Bardossy et al. 1997).

Once the experimental variogram is found, we fit a variogram model. There are several methods such as OLS, WLS, ML and REML. The last two require normality while the first two do not.

To illustrate our methodology, and without loss of generality, we use the ML method. Assuming that $\boldsymbol{Z}_s$ in (6) is a Gaussian process, twice the negative log-likelihood ($\ell_1 = -2 \log L$) is given by

$$\ell_1 \left(\tilde{\boldsymbol{\beta}}, \boldsymbol{\vartheta} \right) = \left(\boldsymbol{Z}_s - \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{\beta}} \right)' \boldsymbol{\Sigma}_{\vartheta}^{-1} \left(\boldsymbol{Z}_s - \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{\beta}} \right) + \log \left(|\boldsymbol{\Sigma}_{\vartheta}| \right) + n \log(2\pi), \tag{10}$$

where $\boldsymbol{\Sigma}_{\vartheta}$ is the covariance matrix of the process $\boldsymbol{Z}_s$, $\widetilde{\boldsymbol{X}} = (\boldsymbol{1}, \boldsymbol{X}_{(k)}) = (\boldsymbol{1}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)$ and $\widetilde{\boldsymbol{\beta}} = (\beta_0, \boldsymbol{\beta}')' = (\beta_0, \beta_1, \ldots, \beta_k)'$. The parameters given in $\boldsymbol{\vartheta}$ usually include the nugget (τ^2), range (ϕ) and partial sill (σ^2), and in the case of the Matérn model, they also include the smoothness parameter κ. The values $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\vartheta}$ are obtained by maximizing iteratively and simultaneously the multivariate normal distribution function, or alternatively by minimizing (10). The minimization is performed in two steps. In the first step, it is assumed that $\boldsymbol{\vartheta}$ is known, and therefore $\boldsymbol{\Sigma}_{\vartheta}$ too. Then the best estimate of the mean parameters $\widetilde{\boldsymbol{\beta}}$ can be obtained by the method of generalized least squares (GLS)

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \left(\widetilde{\boldsymbol{X}}' \boldsymbol{\Sigma}_{\vartheta}^{-1} \widetilde{\boldsymbol{X}} \right)^{-1} \widetilde{\boldsymbol{X}}' \boldsymbol{\Sigma}_{\vartheta}^{-1} \boldsymbol{Z}_s. \tag{11}$$

In the second step, we plug this value into (10), to obtain

$$-2\log L(\boldsymbol{\vartheta}) = \left(\boldsymbol{Z}_s - \widetilde{\boldsymbol{X}}\widehat{\widetilde{\boldsymbol{\beta}}}\right)' \boldsymbol{\Sigma}_{\vartheta}^{-1} \left(\boldsymbol{Z}_s - \widetilde{\boldsymbol{X}}\widehat{\widetilde{\boldsymbol{\beta}}}\right) + \log\left(|\boldsymbol{\Sigma}_{\vartheta}|\right) + n\log(2\pi). \tag{12}$$

Expression (12) is minimized only with respect to $\boldsymbol{\vartheta}$ which makes the process easier and allows to find $\widehat{\boldsymbol{\vartheta}}$. Iterating several times the two previous steps, estimates $\widetilde{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\vartheta}}$ are then found. To start the process, first estimate the trend component by OLS, then obtain the experimental variogram on the residuals. From the experimental variogram, the initial values of the parameter $\boldsymbol{\vartheta}$ in $\boldsymbol{\Sigma}_{\vartheta}$ can be obtained.

We are now ready to perform spatial predictions at a new location ($\boldsymbol{s}_0$) where a set of mixed explanatory variables are observed. We first use the UK method to build the spatial predictions from the DB trend. Thus, assume that on the set of mixed explanatory variables, a new individual $(n+1)$ is observed, with known $\boldsymbol{v}(\boldsymbol{s}_0) = (v_1(\boldsymbol{s}_0), \ldots, v_p(\boldsymbol{s}_0))'$. The distances between the new individual and each of the individuals involved in the model proposed in (2) can be calculated to give $\delta_{0i} = \delta\left(\boldsymbol{v}(\boldsymbol{s}_0), \boldsymbol{v}(\boldsymbol{s}_i)\right)$, $i = 1, \ldots, n$. From these distances, a prediction can be given using a result in Gower (1971) and Cuadras & Arenas (1990), which relates the vector $\boldsymbol{\delta}_0 = \left(\delta_{01}^2, \ldots, \delta_{0n}^2\right)'$ of squared distances with the vector $\boldsymbol{x}\left(\boldsymbol{s}_0\right) = \left(x_1\left(\boldsymbol{s}_0\right), \ldots, x_k\left(\boldsymbol{s}_0\right)\right)'$ of principal coordinates associated to the new individual. This relation is given by

$$\delta_{0i}^2 = \left[\boldsymbol{x}\left(\boldsymbol{s}_0\right) - \boldsymbol{x}\left(\boldsymbol{s}_i\right)\right]' \left[\boldsymbol{x}\left(\boldsymbol{s}_0\right) - \boldsymbol{x}\left(\boldsymbol{s}_i\right)\right]$$

where $\boldsymbol{x}\left(\boldsymbol{s}_i\right) = \left(x_1\left(\boldsymbol{s}_i\right), \ldots, x_k\left(\boldsymbol{s}_i\right)\right)'$, $i = 1, ..., n$. Then, we have that

$$\boldsymbol{x}\left(\boldsymbol{s}_0\right) = \frac{1}{2}\boldsymbol{\Lambda}_{(k)}^{-1}\boldsymbol{X}'_{(k)}(\boldsymbol{b} - \boldsymbol{\delta}_0)$$

where $\boldsymbol{\Lambda}_{(k)}$ is a diagonal matrix containing the k eigenvalues of $\boldsymbol{B}$ associated to $\boldsymbol{X}_{(k)}$, $\boldsymbol{b} = (b_{11}, ..., b_{nn})'$ and $b_{ii} = \boldsymbol{x}'(\boldsymbol{s}_i)\boldsymbol{x}(\boldsymbol{s}_i)$ with $i = 1, ..., n$.

Assume now that we want to predict the value of $Z\left(\boldsymbol{s}_0\right)$ based on a set of observations $\boldsymbol{Z}_s$. The UK predictor is given by

$$\hat{Z}\left(\boldsymbol{s}_0\right) = \sum_{i=1}^{n} \varphi_i Z\left(\boldsymbol{s}_i\right) = \boldsymbol{\varphi}'\boldsymbol{Z}_s$$

where the coefficients vector $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_n)'$ satisfies that $\sum_{i=1}^{n} \varphi_i = 1$. This predictor must satisfy the unbiasedness condition

$$\mathrm{E}\left(\hat{Z}\left(\boldsymbol{s}_0\right) - Z\left(\boldsymbol{s}_0\right)\right) = E\left(\sum_{i=1}^{n} \varphi_i Z(\boldsymbol{s}_i)\right) - \mathrm{E}\left(Z(\boldsymbol{s}_0)\right) = \sum_{i=1}^{n} \varphi_i \mu - \mu = 0$$

Therefore, the following equality is obtained

$$\mathrm{E}\left(\hat{Z}\left(\boldsymbol{s}_0\right)\right) = \mathrm{E}(\boldsymbol{\varphi}'\boldsymbol{Z}_s) = \boldsymbol{\varphi}'\mathrm{E}(\boldsymbol{Z}_s) = \boldsymbol{\varphi}'\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} = \mathrm{E}\left(Z\left(\boldsymbol{s}_0\right)\right) = \widetilde{\boldsymbol{x}}'\left(\boldsymbol{s}_0\right)\widetilde{\boldsymbol{\beta}} \tag{13}$$

where $\widetilde{\boldsymbol{x}}\left(\boldsymbol{s}_0\right) = \left(1, \boldsymbol{x}'\left(\boldsymbol{s}_0\right)\right)' = \left(1, x_1\left(\boldsymbol{s}_0\right), \ldots, x_k\left(\boldsymbol{s}_0\right)\right)'$ is a vector formed by 1 and the principal coordinates of the new individual $\boldsymbol{x}\left(\boldsymbol{s}_0\right)$. From (13) we have that $\boldsymbol{\varphi}'\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{x}}'\left(\boldsymbol{s}_0\right)$. Condition (13) guarantees that the predictor is both unbiased and of minimum variance (Cressie 1993).

The corresponding variance of $\hat{Z}(\boldsymbol{s}_0)$ will be

$$\text{Var}\left(\hat{Z}(\boldsymbol{s}_0)\right) = \boldsymbol{\varphi}'\boldsymbol{\Sigma}_\vartheta\boldsymbol{\varphi} = \sum_{i=1}^{n}\sum_{i'=1}^{n}\varphi_i\varphi_{i'}C\left(e(\boldsymbol{s}_i), e(\boldsymbol{s}_{i'})\right) \geq 0$$

where $\boldsymbol{\Sigma}_\vartheta$ is an $n \times n$ matrix whose (i, i')-th element is $C\left(e(\boldsymbol{s}_i), e(\boldsymbol{s}_{i'})\right)$. This is a condition which describes $\boldsymbol{\Sigma}_\vartheta$ as being positive-definite. If the semivariance function is considered ($\boldsymbol{\Gamma}_\vartheta$) instead of $\boldsymbol{\Sigma}_\vartheta$, the condition is given by

$$-\boldsymbol{\varphi}'\boldsymbol{\Gamma}_\vartheta\boldsymbol{\varphi} = -\sum_{i=1}^{n}\sum_{i'=1}^{n}\varphi_i\varphi_{i'}\gamma_{ii'} \geq 0$$

where $\boldsymbol{\Gamma}_\vartheta$ is an $n \times n$ matrix whose (i, i')-th element is $\gamma_{ii'} = \sigma^2 - C\left(e(\boldsymbol{s}_i), e(\boldsymbol{s}_{i'})\right)$ with $\sigma^2 = C(0, 0)$. In this way, $-\boldsymbol{\Gamma}_\vartheta$ is conditionally positive-definite (Armstrong & Diamond 1984).

On the other hand, the mean-square error of the prediction, $\text{Var}(\boldsymbol{s}_0)$, under the unbiasedness condition, is given by

$$\text{Var}\left(\hat{e}(\boldsymbol{s}_0)\right) = \text{Var}\left[\left(\hat{Z}(\boldsymbol{s}_0) - Z(\boldsymbol{s}_0)\right)\right] = \boldsymbol{\varphi}'\boldsymbol{\Sigma}_\vartheta\boldsymbol{\varphi} + \sigma^2 - 2\boldsymbol{\varphi}'\boldsymbol{c} \tag{14}$$

where $\boldsymbol{c}$ is a vector of dimension n whose i-th element is $C\left(e(\boldsymbol{s}_0), e(\boldsymbol{s}_i)\right)$.

Note now that we must estimate $\boldsymbol{\varphi}$. For this purpose, we replace $\boldsymbol{\Sigma}_\vartheta$ and σ^2 in (14) by their estimates obtained by performing the optimization procedure presented in (11) and (12). Thus, $\boldsymbol{\varphi}$ is found by minimizing the following expression

$$\mathcal{L}(\boldsymbol{\varphi}, \boldsymbol{l}) = \boldsymbol{\varphi}'\boldsymbol{\Sigma}_{\hat{\vartheta}}\boldsymbol{\varphi} + \hat{\sigma}^2 - 2\boldsymbol{\varphi}'\boldsymbol{c} + 2\boldsymbol{l}'\left(\widetilde{\boldsymbol{X}}'\boldsymbol{\varphi} - \widetilde{\boldsymbol{x}}'(\boldsymbol{s}_0)\right)$$

where $\boldsymbol{l}$ is the vector of $(k+1)$ Lagrange multipliers associated with the unbiasedness constraint.

After differentiating with respect to $\boldsymbol{\varphi}$ and $\boldsymbol{l}$, equating the result to zero, and performing some algebraic procedures, the following system is found

$$\begin{pmatrix} \boldsymbol{\Sigma}_{\hat{\vartheta}} & \widetilde{\boldsymbol{X}} \\ \widetilde{\boldsymbol{X}}' & \boldsymbol{0} \end{pmatrix}\begin{pmatrix} \boldsymbol{\varphi} \\ \boldsymbol{l} \end{pmatrix} = \begin{pmatrix} \boldsymbol{c} \\ \boldsymbol{x}(\boldsymbol{s}_0) \end{pmatrix}$$

Solving the system, the coefficient estimates for $\boldsymbol{\varphi}$ and $\boldsymbol{l}$ are given by

$$\begin{aligned} \hat{\boldsymbol{\varphi}}' &= \left[\boldsymbol{c} + \widetilde{\boldsymbol{X}}\left(\widetilde{\boldsymbol{X}}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1}\left(\widetilde{\boldsymbol{x}}(\boldsymbol{s}_0) - \widetilde{\boldsymbol{X}}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\boldsymbol{c}\right)\right]'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1} \\ \hat{\boldsymbol{l}} &= -\left(\widetilde{\boldsymbol{X}}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1}\left(\widetilde{\boldsymbol{x}}(\boldsymbol{s}_0) - \widetilde{\boldsymbol{X}}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\boldsymbol{c}\right) \end{aligned} \tag{15}$$

The estimation of the mean-square prediction error in terms of $\hat{\boldsymbol{\varphi}}$ and $\hat{\boldsymbol{l}}$ can be expressed as

$$\widehat{\text{Var}}\left(\hat{e}(\boldsymbol{s}_0)\right) = \hat{\boldsymbol{\varphi}}'\boldsymbol{c} - \widetilde{\boldsymbol{x}}'(\boldsymbol{s}_0)\hat{\boldsymbol{l}} \quad + \hat{\sigma}^2 - 2\hat{\boldsymbol{\varphi}}'\boldsymbol{c} = \hat{\sigma}^2 - \left(\hat{\boldsymbol{l}}'\widetilde{\boldsymbol{x}}(\boldsymbol{s}_0) + \hat{\boldsymbol{\varphi}}'\boldsymbol{c}\right) \tag{16}$$

Substituting (15) into (16), we finally have

$$\begin{aligned} \widehat{\text{Var}}\left(\hat{e}(\boldsymbol{s}_0)\right) =& \hat{\sigma}^2 - \boldsymbol{c}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\boldsymbol{c} - \left(\widetilde{\boldsymbol{x}}(\boldsymbol{s}_0) - \widetilde{\boldsymbol{X}}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\widetilde{\boldsymbol{X}}\right)'\left(\widetilde{\boldsymbol{X}}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\widetilde{\boldsymbol{X}}\right)^{-1} \\ & \left(\widetilde{\boldsymbol{x}}(\boldsymbol{s}_0) - \widetilde{\boldsymbol{X}}'\boldsymbol{\Sigma}_{\hat{\vartheta}}^{-1}\boldsymbol{c}\right) \end{aligned} \tag{17}$$

For clarity, the procedure can be summarized as follows:

1. Obtain the principal coordinates from the spectral decomposition of the matrix of similarities (or distances) calculated from the explanatory variables.
2. Select those principal coordinates that are most correlated with the regionalized variable $\boldsymbol{Z}_s$. In this step, we recommend using the criterion given in (7) to make a first selection in order to leave out principal coordinates which are poorly correlated with the regionalized variable, and then, using criterion (8) to choose the k most significant principal coordinates using DB regression.
3. Fit the variogram model using OLS, WLS, ML or REML. For ML, estimate $\widetilde{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\hat{\vartheta}}$, which are obtained using iteratively (11) and (12). For OLS and WLS, first construct the residuals $\hat{e}\,(\boldsymbol{s}_i) = Z(\boldsymbol{s}_i) - \hat{Z}(\boldsymbol{s}_i)$, and then, fit the variogram model to the experimental variogram.
4. Make predictions at sample and unsampled points using DBUK (evaluating $\hat{Z}\,(\boldsymbol{s}_0) = \widehat{\boldsymbol{\varphi}}'\boldsymbol{Z}_s$), and calculate prediction error variances using (17).

2.3 Relation with the classical geostatistical model

Model (6) depends on the chosen distance $\delta_{ii'}$. Therefore, when the predictor variables are continuous and the Euclidean distance is used, the distance-based geostatistical model is equivalent with the classical geostatistical model. This equivalence also holds for qualitative and mixture (continuous and categorical) explanatory variables when an appropriate dissimilarity measure is used.

If all the explanatory variables in (3), $\boldsymbol{V}^*$, are continuous, the squared distance is given by (5). Thus, the distance matrix $\boldsymbol{D} = (\delta_{ii'})$ is obtained, and

$$\boldsymbol{A} = -\frac{1}{2}\left[\boldsymbol{f}_{v^*}\mathbf{1}' + \mathbf{1}\boldsymbol{f}'_{v^*} - 2\boldsymbol{V}^*\boldsymbol{V}^{*\prime}\right]$$

where $\boldsymbol{f}_{v^*} = \mathrm{diag}(\boldsymbol{V}^*\boldsymbol{V}^{*\prime})$ is a vector which contains the diagonal terms of the matrix $\boldsymbol{V}^*\boldsymbol{V}^{*\prime}$. Therefore,

$$\boldsymbol{B} = \boldsymbol{HAH} = \boldsymbol{HV}^*\boldsymbol{V}^{*\prime}\boldsymbol{H} = \boldsymbol{VV}' = \boldsymbol{XX}'$$

because $\mathbf{1}'\boldsymbol{H} = \mathbf{0}$. Then, the geostatistical model based on distances introduced in (6) is a centered geostatistical model (3), i.e., it produces the same predictions at p dimensions as the model given in (3).

However, it would not be necessary to consider an Euclidean p-dimensional distance as given in equation (5). Let E_l $(k \leq l \leq n-1)$ be the space spanned by the columns of $\boldsymbol{X}$, where $\boldsymbol{X}$ is a metric scaling solution obtained from a distance applied to the same data. Then, by taking $k > p$, i.e., the most suitable columns of $\boldsymbol{X}$, the distance-based geostatistical model improves the classical geostatistical model when $(\boldsymbol{Z}_s - \widehat{\beta}_0\mathbf{1}) \in E_l$. Note that this is always true for $l = n-1$ with $l > p$.

To illustrate the comparison between the distance-based geostatistical model and the classical geostatistical model, let us focus on R^2. The R_k^2 can be written as

$$R_k^2 = \sum_{j=1}^{k} r^2(\boldsymbol{Z}_s, \boldsymbol{X}_j)$$

Note that $R_k^2 < R_{k+1}^2$ because the principal coordinates $\boldsymbol{X}_j$'s $(j = 1, ..., k)$ are uncorrelated with the principal coordinate $\boldsymbol{X}_{k+1}$. Moreover, R_k^2 is maximized for a given k provided that the first k ordered principal coordinates are selected.

Without loss of generality, suppose $p = 1$ and note that applying the distance-based geostatistical model, R_l^2 can be written as

$$R_l^2 = \sum_{j=1}^{l} r^2(\boldsymbol{Z}_s, \boldsymbol{X}_j), \quad 1 \leq k \leq l \leq n-1$$

because the principal coordinates $(\boldsymbol{X}_1, ..., \boldsymbol{X}_{n-1})$ are uncorrelated. Also, we can represent as $R^2(\boldsymbol{Z}_s, \boldsymbol{V}_1^*)$, the coefficient of determination between $\boldsymbol{Z}_s$ and the explanatory variable $\boldsymbol{V}_1^*$ using the classical geostatistical model. Therefore, if we take $l = k > 1$ (e.g., $k = 2$) then we find that

$$R_l^2 = \sum_{j=1}^{l} r^2(\boldsymbol{Z}_s, \boldsymbol{X}_j) \geq R^2(\boldsymbol{Z}_s, \boldsymbol{V}_1^*)$$

So, the distance-based geostatistical model improves the classical geostatistical model. However, this inequality is not too relevant if $R^2(\boldsymbol{Z}_s, \boldsymbol{V}_1)$ is close to 1, and so, the classical geostatistical model is sufficiently good, and thus the distance-based geostatistical model improvement is not necessary. This result holds when the explanatory variables are qualitative or mixed. This approach aims at improving the predictions by allowing for inclusion of any number of principal coordinates at the sample locations. Note that universal kriging traditionally considers only information of a matrix of size $n \times p$, $\boldsymbol{V}$. Therefore, when we want to improve predictions regardless of the effect that some covariates have on the spatial response variable, the methodology proposed in this paper can be used.

When the explanatory variables in (3) are a mixture or qualitative, if $\boldsymbol{M} = (m_{ii'})$ is the similarity matrix and the selected distance function is $\delta_{ii'}^2 = m_{ii} + m_{i'i'} - 2m_{ii'}$ then, as it was shown for continuous variables, writing $\boldsymbol{A}_q = -\frac{1}{2}[\boldsymbol{m}_v \boldsymbol{1}' + \boldsymbol{1}\boldsymbol{m}_v' - 2\boldsymbol{M}]$, we find that

$$\boldsymbol{B}_q = \boldsymbol{H}\boldsymbol{A}_q\boldsymbol{H} = \boldsymbol{H}\boldsymbol{M}\boldsymbol{H}$$

where $\boldsymbol{m}_v = (m_{11}, m_{22}, .., m_{nn})$ and $\boldsymbol{1}'\boldsymbol{H} = 0$. Therefore, this corresponds to classic multidimensional scaling (Cuadras 1989), and so, the above results are valid.

2.4 Evaluation measures

We consider the root-mean-square prediction error ($RMSPE$) to evaluate the accuracy of the UK and DBUK interpolation methods. The $RMSPE$ is obtained by cross-validation (leave-one-out), which can be used to compare the performance of several interpolation methods. Cross-validation consists of removing one observation from the n sample points (usually related to a neighborhood), and then, with the remaining $(n-1)$ values and the selected variogram model, predict via kriging the variable value at the point location that was removed. This procedure is performed sequentially with each sample point, and thus, a set of n prediction errors is obtained. If the variogram model describes well the spatial autocorrelation structure, then the difference between the observed and predicted values should be small. This procedure is justified because the kriging interpolation methods are accurate, i.e., the predicted and observed values match at the sampled points. In this way, the cross-validation procedure gives an idea of how good the forecasts are, and provides information about which model gives more accurate predictions. Expressions for both the $RMSPE$ and the coefficient of determination (R^2) are given by

$$RMSPE = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{Z}_{[i]}\left(\boldsymbol{s}_i\right) - Z\left(\boldsymbol{s}_i\right)\right)^2}{n}} \tag{18}$$

$$R^2 = 1 - \sum_{i=1}^{n}\left(\hat{Z}_{[i]}\left(\boldsymbol{s}_i\right) - Z\left(\boldsymbol{s}_i\right)\right)^2 \Big/ \sum_{i=1}^{n}\left(Z\left(\boldsymbol{s}_i\right) - \bar{Z}\left(\boldsymbol{s}_i\right)\right)^2 \tag{19}$$

where $\hat{Z}_{[i]}\left(\boldsymbol{s}_i\right)$ is the predicted value obtained from cross-validation, and $Z\left(\boldsymbol{s}_i\right)$ is the sampled value at location $\boldsymbol{s}_i$.

A variation of the previous methodology is given by dividing the sample into two sub-samples; the first sub-sample is used for variogram modeling, and the other sub-sample is used to validate the kriging method. After that, validation measures can be constructed from the observed and predicted values (Bivand et al. 2008). If the interpolation method is adequate, the $RMSPE$ should be as small as possible (close to zero) and R^2 should be close to 1.

3 Simulated experiment

To evaluate our proposal, a simulation study was performed under certain scenarios. In these scenarios, for the classical UK method we used the mixed explanatory variables in the traditional form, and for the DBUK method we used the principal coordinates obtained from the mixed explanatory variables. In the latter case, the principal coordinates were obtained using the criterion given in (7) to make a first selection, leaving out those principal coordinates with correlation close to zero. Then, we used the criterion given in (8) to choose the most significant principal coordinates in the DB regression.

The trend was built considering a nominal variable associated with three fixed regions in the unit square (see Figure 1). Since there are three regions, only two dummy variables ($\boldsymbol{D}_2$, $\boldsymbol{D}_3$) were considered to avoid singularity problems. In addition, we considered a dichotomous random variable $\boldsymbol{V}_1 = (v_1(\boldsymbol{s}_1), \ldots, v_1(\boldsymbol{s}_n))'$, where $v_{i1} = v_1(\boldsymbol{s}_i) \sim Bernoulli(p = 0.4)$ for $i = 1, \ldots, n$ with $n = 50, 100, 150$, and assumed that the error $\varepsilon(\boldsymbol{s}_i)$ follows a zero-mean stationary isotropic Gaussian process with a covariance function generated from a specific variogram model with a set of values for the following parameters: nugget (τ^2), (practical) range (ϕ), partial sill (σ^2), and sample size (n). Four theoretical variogram models were used Exponential, Stable ($\kappa = 1.5$), Gaussian and Spherical, and these were adjusted by the maximum likelihood method. The simulated scenarios are presented in Table 1. The trend parameters were fixed as $\beta_0 = 10, \beta_1 = -4, \beta_2 = 2$ and $\beta_3 = -4$ associated to the dichotomous variable $\boldsymbol{V}_1$ and the spatial coordinates w_{x_i} and w_{y_i}, $i = 1, \ldots, n$ respectively. Thus, the simulated regionalized process is given by

$$Z(\boldsymbol{s}_i) = \beta_0 + \beta_1 v_{i1} + \beta_2 D_{i2} w_{x_i} + \beta_3 D_{i3} w_{y_i} + \varepsilon(\boldsymbol{s}_i) \tag{20}$$

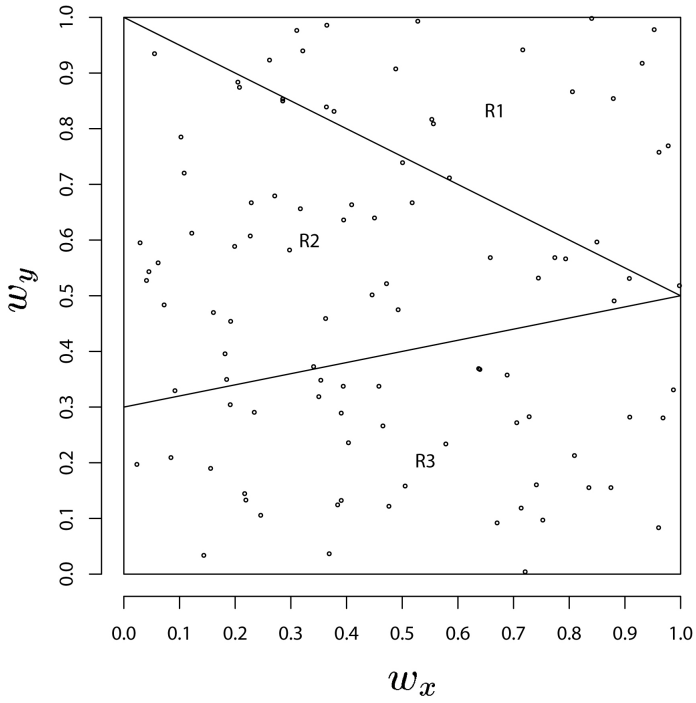


Fig. 1 Location of sampling points and associated regions defining the nominal variable

We followed the methodology and steps described in Section 2.2. To examine the versatility of the DBUK method with respect to the classical UK method, only ten principal coordinates were considered for the purposes of a homogeneous comparison, 96 scenarios were simulated (as shown in Table 1), and for each of them, the whole process was repeated 100 times.

Table 1 Simulated scenarios identified by the corresponding numbers

Model parameters				Variogram models			
τ^2	σ^2	ϕ	n	Exponential	Stable	Gaussian	Spherical
0	0.5	0.15	50	1	25	49	73
			100	2	26	50	74
			150	3	27	51	75
		0.40	50	4	28	52	76
			100	5	29	53	77
			150	6	30	54	78
	1	0.15	50	7	31	55	79
			100	8	32	56	80
			150	9	33	57	81
		0.40	50	10	34	58	82
			100	11	35	59	83
			150	12	36	60	84
0.5	0	0.15	50	13	37	61	85
			100	14	38	62	86
			150	15	39	63	87
		0.40	50	16	40	64	88
			100	17	41	65	89
			150	18	42	66	90
	1	0.15	50	19	43	67	91
			100	20	44	68	92
			150	21	45	69	93
		0.40	50	22	46	70	94
			100	23	47	71	95
			150	24	48	72	96

Under each scenario and simulation, the various theoretical models were fitted to the experimental variogram by maximum likelihood, and the $RMSPE$ statistic given in (18) was calculated by leave-one-out cross-validation (LOOCV). The results in terms of the $RMSPE$ are shown in Table 2. If we select, for example, the scenario number 60 (see Table 1) corresponding to a Gaussian variogram model with parameters $\tau^2 = 0$, $\sigma^2 = 1$, $\phi = 0.40$ and $n = 150$, the average $RMSPE$ was lower for DBUK (0.14) than for UK (0.18). In general, the same behavior was found in the other scenarios. Therefore, there was a significant gain in error reduction using the proposed distance-based method. The greatest gain was obtained for the Gaussian model because there was a reduction of 4.3% on average, while the reduction for the exponential model was 3.6%. The reduction for the spherical and stable models was of 3.0% and 3.3%, respectively.

Table 2 Average $RMSPE$ under the UK and DBUK methods for the scenarios presented in Table 1

Model parameters				Variogram models							
				Exponential		Stable $\kappa = 1.5$		Gaussian		Spherical	
τ^2	σ^2	ϕ	n	UK	DBUK	UK	DBUK	UK	DBUK	UK	DBUK
0	0.5	0.15	50	0.73	0.70	0.64	0.63	0.64	0.61	0.70	0.68
			100	0.64	0.61	0.58	0.55	0.46	0.43	0.58	0.55
			150	0.61	0.59	0.56	0.55	0.35	0.32	0.52	0.51
		0.40	50	0.59	0.56	0.56	0.55	0.29	0.30	0.47	0.46
			100	0.48	0.46	0.49	0.47	0.20	0.12	0.37	0.35
			150	0.43	0.43	0.47	0.47	0.16	0.13	0.33	0.32
	1	0.15	50	1.00	0.97	0.89	0.85	0.87	0.84	0.96	0.94
			100	0.89	0.85	0.80	0.77	0.62	0.60	0.80	0.77
			150	0.85	0.83	0.78	0.76	0.47	0.44	0.72	0.71
		0.40	50	0.81	0.77	0.77	0.75	0.33	0.38	0.64	0.64
			100	0.66	0.64	0.68	0.65	0.22	0.13	0.50	0.50
			150	0.60	0.59	0.65	0.65	0.18	0.14	0.45	0.44
0.5	0.5	0.15	50	1.06	1.05	1.01	0.97	1.01	0.97	1.04	1.00
			100	1.00	0.94	0.95	0.91	0.93	0.89	0.97	0.93
			150	0.98	0.95	0.94	0.93	0.90	0.88	0.95	0.92
		0.40	50	0.99	0.97	0.96	0.92	0.88	0.84	0.93	0.90
			100	0.92	0.88	0.91	0.86	0.81	0.79	0.87	0.84
			150	0.90	0.87	0.90	0.87	0.79	0.78	0.85	0.83
	1	0.15	50	1.27	1.22	1.18	1.14	1.18	1.14	1.24	1.18
			100	1.18	1.13	1.11	1.06	1.05	1.02	1.13	1.09
			150	1.15	1.13	1.09	1.07	0.99	0.97	1.09	1.06
		0.40	50	1.14	1.09	1.09	1.06	0.93	0.91	1.04	1.01
			100	1.03	0.99	1.02	0.98	0.84	0.83	0.94	0.92
			150	1.00	0.97	1.01	0.98	0.81	0.80	0.91	0.90

According to Table 2, the DBUK method is generally better than the UK method because under $\tau^2 = 0$ and $\tau^2 = 0.5$, the averages of $RMSPE$ were 3.8% and 3.3%, respectively, higher in the UK than in the DBUK method. Additionally, the average $RMSPE$ under the UK and DBUK methods decreases as the sample size increases. In all cases, the RMSPE was smaller for $\sigma^2 = 0.4$, as compared to $\sigma^2 = 0.15$ independently of the method.

4 Applications

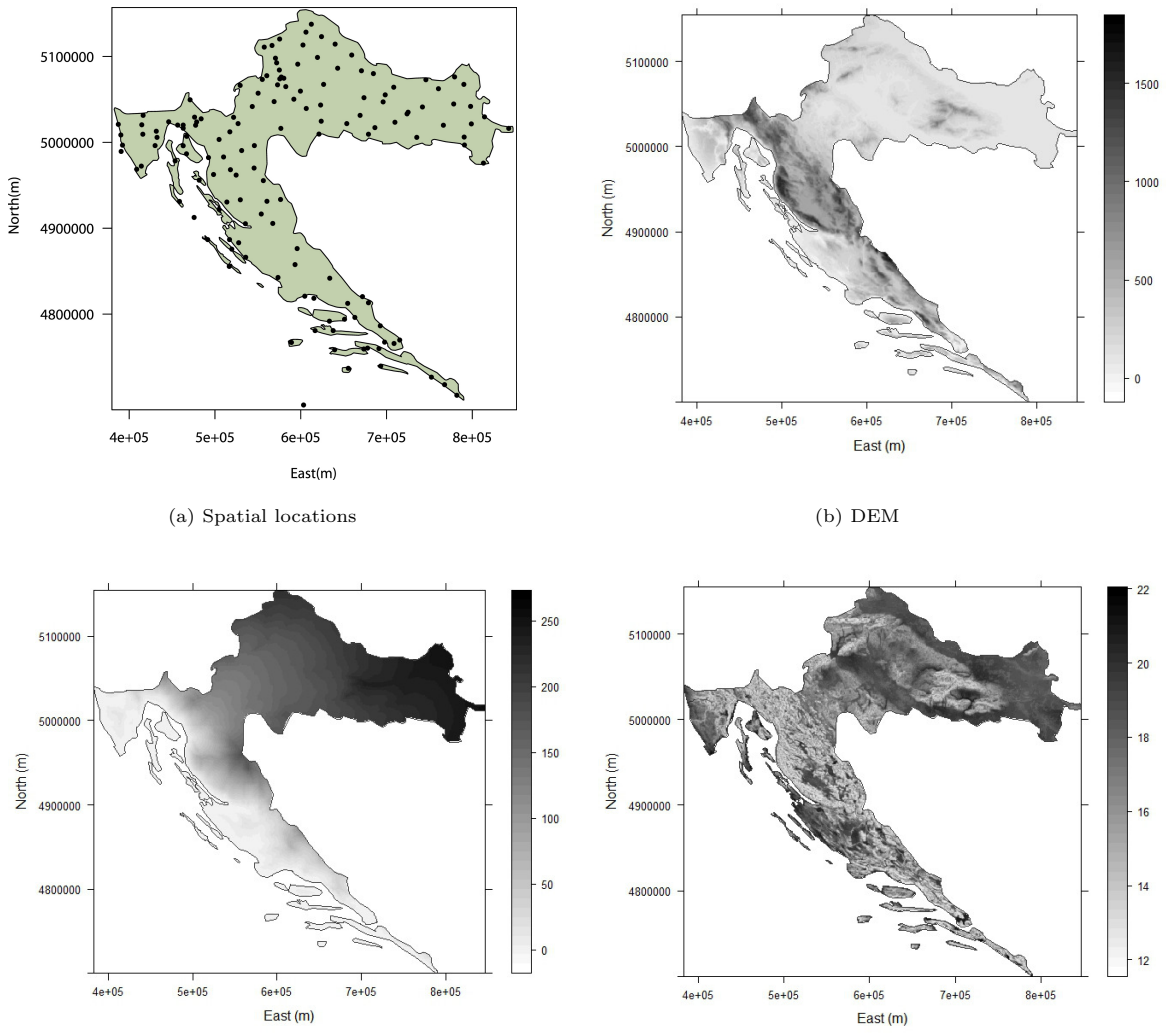
In the two considered applications, we fitted two models for the mean of $\boldsymbol{Z}_s$, the first one using (2) with original mixed explanatory variables for the classical UK, and the second one using (6) with the principal coordinates generated from the mixed explanatory variables for the DBUK method. In the latter case, we used the criterion (7) to make a first selection, removing those principal coordinates with correlation close to zero. We then used criterion (8) to choose the most significant principal coordinates in the DB regression. In the first application, the explanatory variables are continuous, while in the second, these are mixed.

For both methods, the residuals obtained from (2) and (6) were used to obtain the experimental variograms (classical and robust) and their associated theoretical variograms. This procedure was performed in different directions to evaluate the isotropy, and in each case, a theoretical model (spherical, exponential, Gaussian and Stable), $\gamma(\boldsymbol{h}, \boldsymbol{\vartheta})$, compatible with the experimental variogram was chosen. We estimated the parameters $\hat{\boldsymbol{\vartheta}}$ using the classical methods of OLS, WLS, ML or REML. We used the fitted model of variogram, $\hat{\gamma}(\boldsymbol{h}; \hat{\boldsymbol{\vartheta}})$, for interpolation purposes under both the UK and DBUK methods. The final step was prediction at sampled and unsampled points.

4.1 Average daily temperature in Croatia

The average daily temperature in Croatia was measured on January 25, 2008 at 154 meteorological stations. This information is taken from http://spatial-analyst.net/book/HRclim2008, and was provided by Melita Perčec Tadić, Croatian Meteorological Organization and Hydrological Services (Hengl 2009). Croatia is a relatively small country, but has several different climates, which are a result of its specific position on the Adriatic Sea and fairly diverse topography ranging from plains on the east, through a central mountainous region separating the continental from the maritime part of the country. The study region is characterized by a wide range of topographical and climatological features, which allows to properly assess the proposed methodology with respect to the traditional one, because the earth's average temperatures in this region are strongly influenced by the topography. Temperature measurements are automatically collected at 154 meteorological stations. The spatial distribution of the stations is not optimal (Zaninovic et al. 2008), in the sense that there is a certain under-sampling at higher elevations and in areas with lower population density (for practical reasons, the areas of higher population density have been given a priority). Consequently, the accuracy of the mapping will be lower at higher elevations and highlands (Hengl 2009, Perčec Tadić 2010).

Figure 2 shows: (a) the spatial locations (w_x, w_y) of 154 meteorological stations, (b) the Digital Elevation Model (DEM, in meters), (c) the weighted topographic distance from the coastline (DSEA, in km), (d) the Topographic Wetness Index (TWI). For spatial interpolation, the spatial locations were standardized to give equal weight to all dimensions.



(a) Spatial locations

(b) DEM

(c) DSEA

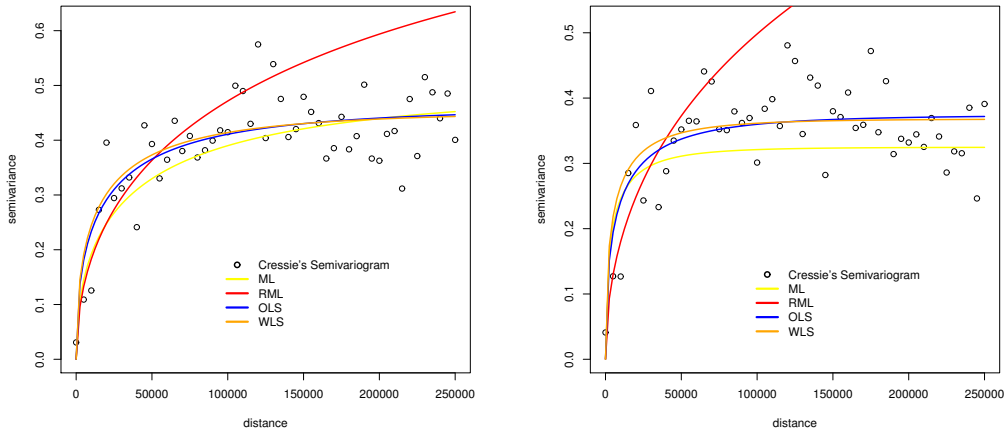
(d) TWI

Fig. 2 (a) Spatial locations of meteorological stations in Croatia, and static topographic predictors: (b) Digital Elevation Model (DEM, in meters), (c) topographically weighted distance from the coast line (DSEA, in km) and (d) Topographic Wetness Index (TWI)

The geographical coordinates (latitude and longitude) were transformed to a Cartesian coordinate system (w_x and w_y), the datum used was WGS84 zone 33, and the transformation method used is known as Bursa Wolf. The principal coordinates were calculated from the spectral decomposition generated by

the spatial coordinates, w_x and w_y. The first 38 principal coordinates were significant at 5% level using criterion (8), and therefore, these were the most related coordinates with the earth's average temperature. However, the fit was made only with the first four for simplicity. In parallel, for the classical approach, a regression was carried out using w_x, w_y, DEM, TWI and DSEA (w_x and TWI were not statistically significant), considering a linear structure of order one. Subsequently, a fitted variogram was built from the residuals obtained from both the classical and DB methods (in the classical case the residuals were anisotropic[1]).

For the two methods, the Matérn variogram models ($\kappa = 1.5$), exponential, spherical and stable (with $\kappa = 0.5$) were considered. The parameters were estimated using the procedures OLS, WLS [2], ML and REML. The exponential model provide the best fit for both the classical and DB methods. This model presented lowest weighted sum of squares (WSS). The resulting fitted variograms are shown in Figure 3, and Table 3 reports the values of the parameters obtained with their corresponding WSS.



(a) Linear regression for the classical method

(b) Linear regression for the DB method

Fig. 3 Fitted robust experimental variogram, from an exponential model estimated with OLS, WLS, ML and REML

Once the variograms were fitted, the predicted maps of the average temperature and their corresponding standard errors for the kriging prediction were obtained (see Figures 4 and 5). In the DBUK case, the principal coordinates of sampled and additional points generated for the original coordinate grid (w_x and w_y) were considered. The predictions are presented in Figure 4 indicating a high agreement between UK and DBUK methods. Figure 5 shows prediction standard errors maps for both the UK and

[1] The intamap library of R was used (Pebesma et al. 2010), which gave an angle of -70.11126 and a rate of 1.141188.

[2] $|N(h_j)|$ was considered for the weights.

Table 3 Comparison between the DB and the classical method with fitted parameter values for an exponential variogram minimizing the weighted sum of squares (WSS)

(a) DB method

Number of principal coordinates with detrending DB	τ^2	σ^2	ϕ	κ	WSS
4	0.000	0.365	13981.052	0.5	22.112

(b) Classical method

Classic parameter for detrending	τ^2	σ^2	ϕ	κ	WSS
DEM,DSEA,w_y	0.124	0.312	30973.727	0.5	26.601

DBUK methods. In general, the prediction standard errors for DBUK are smaller throughout the region of study.

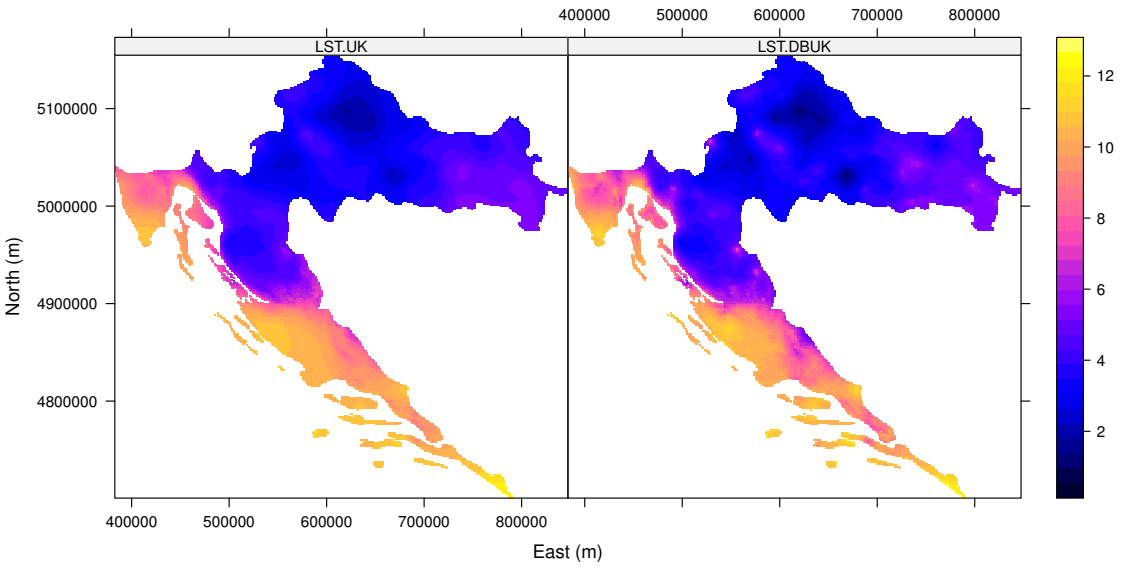


Fig. 4 Predicted maps of average daily Land Surface Temperature (LST) in Croatia

Finally, to assess the practical advantages of DBUK over the UK method, the LOOCV was performed and the RMSPE and R^2 statistics were generated from the geospt library (Melo et al. 2012). The results obtained are shown in Table 4. In the DBUK case, 4 and 38 principal coordinates were considered. However, although expanding the number of principal coordinates improves the predictions, it is not advisable to leave many principals coordinates since it generates more calculation time. So, the DBUK model is recommended using a moderate number of principal coordinates in order to obtain good predictions.

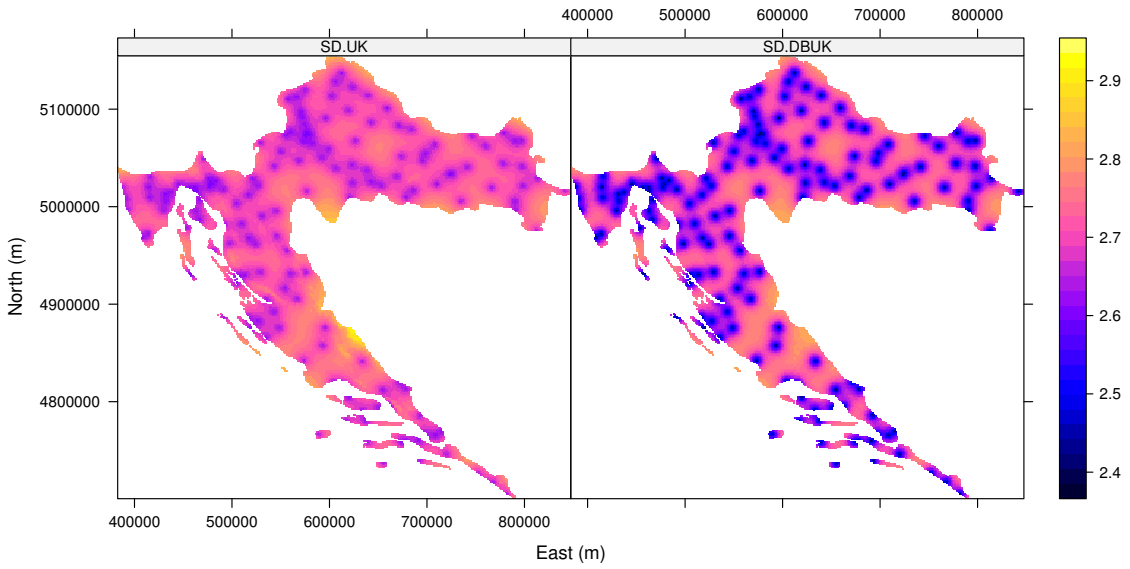


Fig. 5 Prediction standard errors maps for the average daily LST in Croatia

Table 4 Comparison between UK and DBUK for the average daily temperature in Croatia using LOOCV

	UK	DBUK (4)	DBUK (38)
RMSPE	0.555	0.523	0.381
R^2	0.686	0.722	0.851

In this case, DBUK method shows a clear superiority over the classical UK, when increasing the principal coordinates. This fact can be seen in the next application.

4.2 Calcium content

This data set considered soil samples collected with a Dutch type drill on an incomplete regular lattice at a spacing of approximately 50 meters, with geographic coordinates, north and east 900 meters apart in both directions. The soil samples were taken from a 0-20 cm depth layer at each of the 178 locations (Figure 6). Magnesium and calcium content were measured in $mmol_c/dm^3$. In this application, we considered only the calcium content. The study region was divided into three sub-regions (soil type) which experienced different soil management regimes. This characterization motivates the application of our proposed method.

This data set had information about the calcium content, the spatial coordinates (w_x, w_y), altitude, and sub-region code of each sample, which is associated with three periods of fertilization in different places or areas. The data are taken from Capeche et al. (1997), and the main objective of the study

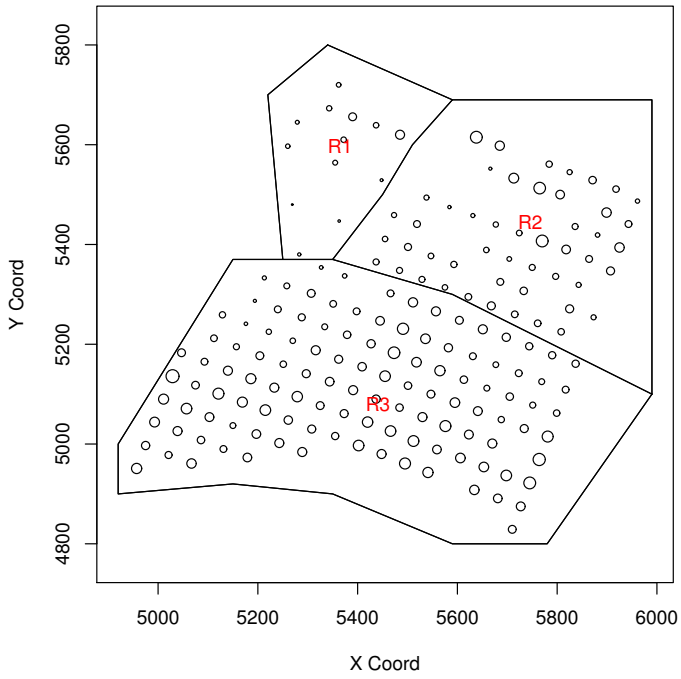


Fig. 6 Circle plot of calcium content with lines delimiting sub-regions (sampling locations). The sizes of the circles are proportional to the calcium content

was adequate land use planning that allows rational and sustainable management, avoiding the erosion process. This is important in order to allocate subsidies in the experimental fields to perform searches that will be extrapolated to similar soils and climatic zones.

Table 5 contains the fitted parameter values using a spherical variogram, and two times the log-likelihood ($2 \log L$) for both the proposed DB method and the classical one. In the DB case, the principal coordinates were built using the spatial coordinates, altitude, and the nominal variable defining the sub-region. The results presented Table 5(a) show a steady increase in $2 \log L$ when increasing the number of principal coordinates. There is a significant gain as $2 \log L$ goes from -1272.03 to -1189.96 when the number of principal coordinates goes from 0 to 15. Table 5(b) shows the values of $2 \log L$ for the classical case. Note that $2 \log L$ increases when considering soil type, but there is no gain by adding the spatial coordinates or altitude. To compare the two methods, 15 principal coordinates in the DB method were considered (see Table 5(a)) for the detrending because the other principal coordinates were not significant at 5% level using criterion (8). The explanatory variables (soil type, altitude and linear spatial trend) were taken in the classical method (see Table 5(b)). Once obtained the variogram models, the UK and DBUK methods were implemented, and the prediction maps are shown in Figure 7.

Table 5 Comparison between the DB and the classical method with fitted parameter values for a spherical variogram using maximum likelihood

(a) DB method

Number of principal coordinates with detrending DB	τ^2	σ^2	ϕ	$2\log L$
α_0	23.23	111.69	244.90	-1272.03
2	0	92.68	112.20	-1265.17
4	0	79.98	102.28	-1251.60
6	0	67.78	91.52	-1234.62
15	0	49.78	80.571	-1189.96

(b) Classical method

Classic parameter for detrending	τ^2	σ^2	ϕ	$2\log L$
β_0	23.23	111.69	244.90	-1272.03
soil type	0	93.00	111.97	-1266.09
soil type, altitude	0	92.69	111.53	-1266.06
soil type, linear spatial trend	0	87.53	107.45	-1261.18
soil type, altitude, linear spatial trend	0	84.52	104.09	-1259.17

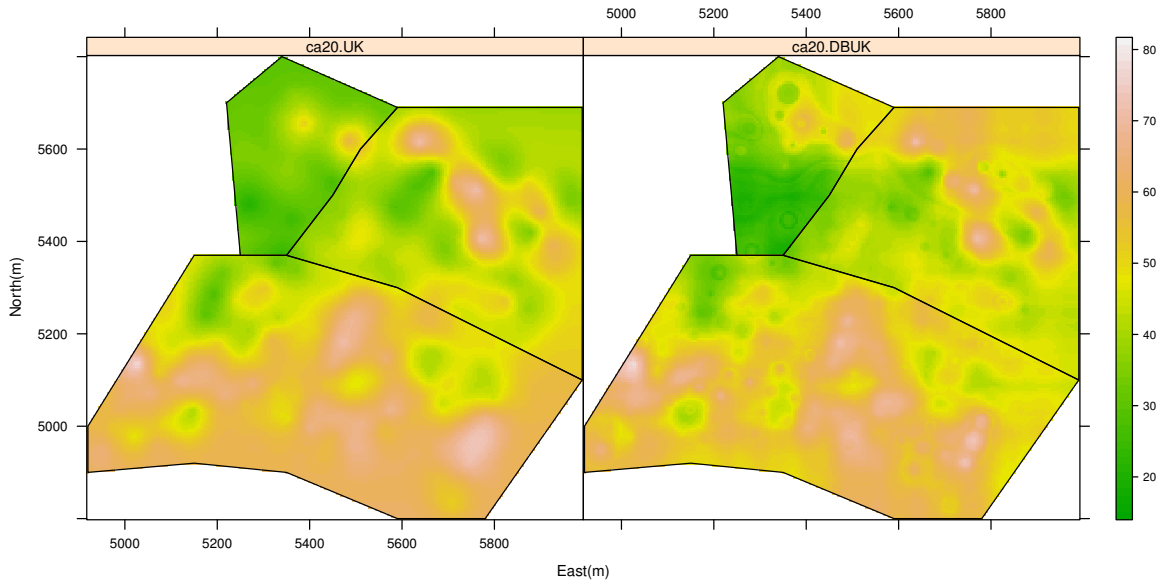


Fig. 7 Predicted maps of soil calcium content including sub-region

The standard errors of the two analyzed methods are shown in Figure 8. There was a clear reduction in the standard errors shown by the DBUK method with respect to the UK method. LOOCV results are shown in Table 6, where it is highlighted a gain of about 10% of the proposed DBUK method over the classical UK one.

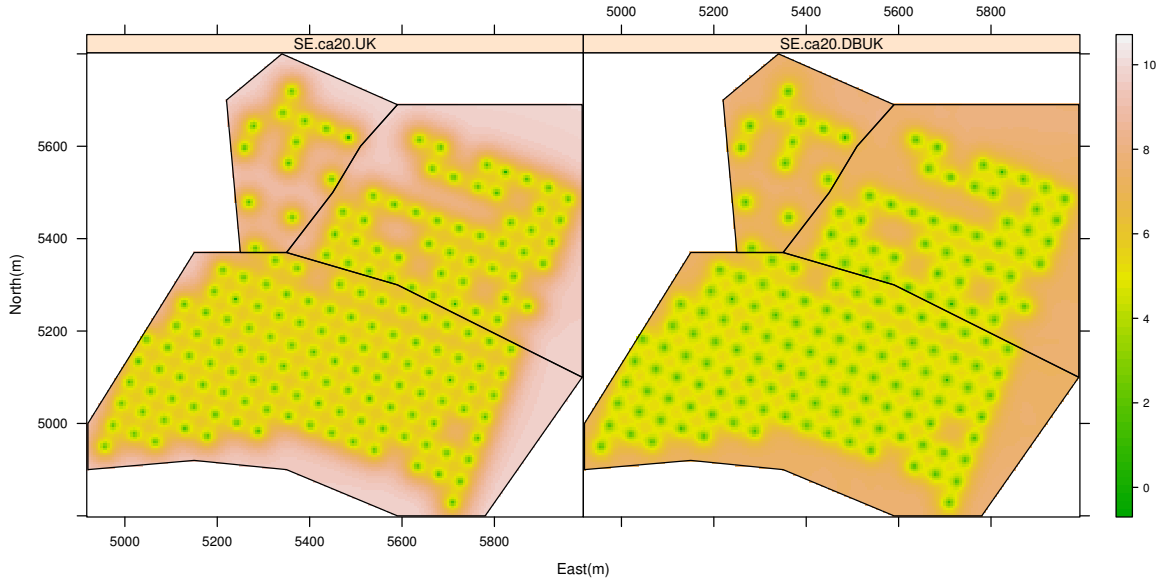


Fig. 8 Maps of prediction standard errors for soil calcium content including the sub-regions

Table 6 Comparison between UK and DBUK for the calcium content using cross-validation

	UK	DBUK (4)	DBUK (15)
RMSPE	7.913	7.734	7.101
R^2	0.487	0.510	0.587

5 Conclusions

We proposed a new DB method for spatial prediction. An important feature of our method is that it provides a framework for dealing with different types of explanatory variables (continuous, discrete, and categorical) when performing spectral analysis. However, if all explanatory variables were continuous or categorical, the DB method can also be used. We have shown that the DBUK performs better than the classical UK, but for inference, the interpretability of the relation between covariates and the response is lost. Thus, if we want to improve the spatial predictions, the DBUK may be employed when the interest is not on the relationship between explanatory variables and the spatial response.

Modeling uncertainty for regionalized variables as a function of mixed variables (continuous, binary or categorical) can be relevant in many disciplines of geosciences and environmental areas. The proposed method shows, both through simulations and applications, some advantages in reducing the error with respect to the classical UK method. This error reduction is associated with a better modeling of the trend, and a smaller error in fitting and modeling the variogram. Among many other possible causes, the error is generated by omission of variables and by considering incorrect functional forms. In general,

by increasing the number of principal coordinates, the DBUK method produces good estimates of the regionalized variable. Then, this method is very flexible and gives good results in practice.

In a variety of studies, it is practically impossible to detect variability among areas. Thus, the proposed DBUK, which takes full advantage of the existing information, is expected to be useful in these cases. Although the correlation between an explanatory variable and the response variable can be low, the key point in the proposed method is that it takes into account the correlation between the principal coordinates (built with the existing explanatory variables) and the response variable. The DBUK method can produce better estimates of the regionalized variable if the number of principal coordinates is increased. This is possible by including more significant principal coordinates in the trend and in the spatial variogram model; the second application illustrates this fact. However, a simulation study should be performed and included for completeness.

Acknowledgements Work partially funded and supported by: grant MTM2016-78917-R from the Spanish Ministry of Science and Education, Core Spatial Data Research (Faculty of Engineering, COL0013969, Universidad Distrital Francisco José de Caldas), and Applied Statistics in Experimental Research, Industry and Biotechnology (COL0004469, Universidad Nacional de Colombia).The authors are grateful to the AE and the referees for their helpful comments and suggestions that have improved the manuscript.

References

Armstrong, M. & Diamond, P. (1984), 'Testing variograms for positive-definiteness', *Mathematical Geology* **16**, 407–421.

Bardossy, A., Haberlandt, U. & Grimm-Strele, J. (1997), Interpolation of groundwater quality parameters using additional information, Technical report, In GeoENV I (Geostatistics for Environmental Applications), 189-200, Kluwer Academic Publ., Dordrecht.

Bivand, R., Pebesma, E. & Rubio, V. (2008), *Applied Spatial Data Analysis with R*, Springer, New York.

Capeche, C. L., Macedo, J. R., Manzatto, H. R. H. & Silva, E. F. (1997), Caracterização pedológica da fazenda angra - pesagro/rio, Technical report, Technical Report: Estação experimental de Campos (RJ). In Informação, globalização, uso do solo, Rio de Janeiro.

Cressie, N. (1993), *Statistics for Spatial Data*, Revised Edition. John Wiley & Sons Inc., New York.

Cressie, N. & Hawkins, D. M. (1980), 'Robust estimation of the variogram: I', *Mathematical Geology* **12**(2), 115–125.

Cuadras, C. M., Arenas, C. & Fortiana, J. (1996), 'Some computational aspects of a distance-based model for prediction', *Communications in Statistics - Simulation and Computation* **25**(3), 593–609.

Cuadras, C. M. (1989), Distance analysis in discrimination and classification using both continuous and categorical variables., *in* Y. Dodge, ed., 'Statistical Data Analysis and Inference', Amsterdam: North-Holland Publishing Co., pp. 459–473.

Cuadras, C. M. (1993), 'Interpreting an inequality in multiple regression', *The American Statistician* **47**(4), 256–258.

Cuadras, C. M. & Arenas, C. (1990), 'A distance based regression model for prediction with mixed data', *Communications in Statistics A - Theory and Methods* **19**, 2261–2279.

Gower, J. (1968), 'Adding a point to vector diagrams in multivariate analysis', *Biometrika* **55**, 582–585.

Gower, J. (1971), 'A general coefficient of similarity and some of its properties', *Biometrics* **27**, 857–871.

Hengl, T. (2009), *A Practical Guide to Geostatistical Mapping*, 2nd edn, University of Amsterdam, Amsterdam.

Jin, R., Chen, W. & Simpson, T. (2001), 'Comparative studies of metamodeling techniques under multiple modeling criteria', *Journal of Structural & Multidisciplinary Optimization* **23**, 1–13.

Joseph, V. R., Hung, Y. & Sudjianto, A. (2008), 'Blind kriging: A new method for developing metamodels.', *Journal of Mechanical Design* **3**, 31–102.

Le, N. & Zidek, J. (2006), *Statistical Analysis of Environmental Space-Time Processes*, Springer.

Mardia, K. V., Kent, J. T. & Bibby, J. M. (2002), *Multivariate Analysis*, Academic Press, Inc, London.

Melo, C., Santacruz, A. & Melo, O. (2012), *geospt: An R package for spatial statistics.* R package version 1.0-0.

URL: *geospt.r-forge.r-project.org/*

Pebesma, E.and Cornford, D., Dubois, G., Heuvelink, G. B. M., Hristopoulos, D., Pilz, J., Stoehlker, U., Morin, G. & Skoien, J. O. (2010), 'Intamap: the design and implementation of an interoperable automated interpolation web service', *Computers & Geosciences* **37**, 343–352.

URL: *http://dx.doi.org/10.1016/j.cageo.2010.03.019*

Perčec Tadić, M. (2010), 'Gridded croatian climatology for 1961-1990', *Theoretical and Applied Climatology* pp. 1434–4483.

Sacks, J., Welch, W., Mitchell, T. J. & Wynn, H. P. (1989), 'Design and analysis of computer experiments', *Statistical Science* **4**, 409–423.

Santner, T., Williams, B. & Notz, W. (2003), *The Design and Analysis of Computer Experiments*, Springer-Verlag, New York.

Sokal, R. R. & Michener, C. D. (1958), 'A statistical methods for evaluating relationships', *University of Kansas Science Bulletin* **38**, 1409–1448.

van de Kassteele, J., Stein, A., Dekkers, A. & Velders, G. (2009), 'External drift kriging of NOx; concentrations with dispersion model output in a reduced air quality monitoring network', *Environmental and Ecological Statistics* **16**, 321–339.

Wackernagel, H. (2003), *Multivariate Geostatistics: An Introduction with applications*, Third Completely Revised Edition. Springer-Verlag, New York.

Zaninovic, K., Gajic-Capka, M. & Percec-Tadic, M. (2008), Klimatski atlas Hrvatske, climate altas of Croatia 1961, 1990, 1971 2000., Technical report, Meteorological and Hydrological Service Republic of Croatia, Zagreb.