Geoderma xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Geoderma

journal homepage: www.elsevier.com/locate/geoderma

Accounting for non-stationary variance in geostatistical mapping of soil properties

Alexandre M.J-C. Wadoux[a,*], Dick J. Brus[b], Gerard B.M. Heuvelink[a]

[a] Soil Geography and Landscape group, Wageningen University, The Netherlands
[b] Biometris, Wageningen University & Research, The Netherlands

ARTICLE INFO

Keywords:
Geostatistics
Pedometrics
Kriging
Non-stationarity
Uncertainty assessment
REML

ABSTRACT

Simple and ordinary kriging assume a constant mean and variance of the soil variable of interest. This assumption is often implausible because the mean and/or variance are linked to terrain attributes, parent material or other soil forming factors. In kriging with external drift (KED) non-stationarity in the mean is accounted for by modelling it as a linear combination of covariates. In this study, we applied an extension of KED that also accounts for non-stationary variance. Similar to the mean, the variance is modelled as a linear combination of covariates. The set of covariates for the mean may differ from the set for the variance. The best combinations of covariates for the mean and variance are selected using Akaike's information criterion. Model parameters of the selected model are then estimated by differential evolution using the Restricted Maximum Likelihood (REML) in the objective function. The methodology was tested in a small area of the Hunter Valley, NSW Australia, where samples from a fine grid with gamma K measurements were treated as measurements of the variable of interest. Terrain attributes were used as covariates. Both a non-stationary variance and a stationary variance model were calibrated. The mean squared prediction errors of the two models were somewhat comparable. However, the uncertainty about the predictions was much better quantified by the non-stationary variance model, as indicated by the mean and median of the standardized squared prediction error and by accuracy plots. We conclude that the non-stationary variance model is more flexible and better suited for uncertainty quantification of a mapped soil property. However, parameter estimation of the non-stationary variance model requires more attention due to possible singularity of the covariance matrix.

1. Introduction

Standard geostatistical mapping approaches predict a soil variable of interest at the unsampled nodes of a fine grid using measurements of this variable at sampling locations. In many cases predictions can be improved by exploiting a relation between the soil variable and one or more environmental covariates of which maps are available, such as terrain attributes derived from a digital elevation model and remote sensing images. This is usually done by modelling the soil variable as the sum of a linear combination of covariates and a spatially autocorrelated residual. This leads to Kriging with External Drift (KED) (Goovaerts, 1997). In situations where the covariates explain a considerable part of the variation of the soil variable, KED is superior to simple or ordinary kriging that both assume that the mean of the soil variable is constant within a global or local neighbourhood and not dependent on covariates.

In KED we allow for a non-stationary mean, but the variance is assumed stationary (i.e., constant). More specifically, it is assumed that the covariance between the soil variable Z at two locations $\mathbf{s}$ and $\mathbf{s} + \mathbf{h}$ only depends on the separation vector $\mathbf{h}$: $Cov(Z(\mathbf{s}),Z(\mathbf{s} + \mathbf{h})) = C(\mathbf{h})$. Taking $\mathbf{h} = \mathbf{0}$ shows that the variance is assumed constant: $Var(Z(\mathbf{s})) = C(\mathbf{0})$ for all $\mathbf{s}$. However, in many cases the assumption of a stationary variance may be implausible, i.e. when the residual spatial variation is substantially different in different parts of the study area. For instance, McBratney and Webster (1981) identified several discontinuities in the variograms of soil colour and pH along a transect in north-east Scotland. The authors attributed the changes to boundaries between soil types. Similarly, Voltz and Webster (1990) found important differences between topsoil clay content variograms of contrasting Jurassic sediments.

In some cases, non-stationarity in the variance can be solved by transforming the data prior to geostatistical modelling, e.g. by a square-root or log-transformation (e.g., Jacques et al., 1999). Several solutions have been proposed in case a transformation does not solve the problem. Pintore and Holmes (2004) and later Haskard and Lark (2009) proposed to account for a non-stationary variance by spectral

* Corresponding author at: Soil Geography and Landscape group, Wageningen University, Droevendaalsesteeg 3, Wageningen 6708 BP, The Netherlands.
E-mail address: alexandre.wadoux@wur.nl (A. M.J-C. Wadoux).

https://doi.org/10.1016/j.geoderma.2018.03.010
Received 17 November 2017; Received in revised form 21 February 2018; Accepted 12 March 2018

tempering. The method tempers a spectrum based on a stationary correlation matrix, but the modelled covariance structure can vary spatially while maintaining positive-definiteness. The authors showed that modifying the spectrum of the data according to a covariate on a transect gave a more realistic variance model for a case study on rates of emission of nitrous oxide from soils. Alternatively, McBratney and Minasny (2013) proposed to equalize variogram parameters by deformation of the geographic space. This method renders a stationary covariance function in the transformed space. Spatial predictions made in the transformed space are then back-transformed to the original geographic space. However, while this approach addresses differences in spatial correlation, it does not solve the non-stationary variance problem.

The work presented here builds on the work of Lark (2009) and Marchant et al. (2009). They demonstrated how a model in which the variance is a function of the spatial coordinate or covariates can be fitted by REML, and how such model can be used in geostatistical prediction of soil properties. The same approach is applied by Brus et al. (2016) in three-dimensional soil property mapping. They assumed that the residual variance is a stepwise or continuous function of depth, while in the horizontal plane, at a given depth, the residual variance was assumed constant.

The objective of this study is to test the approach proposed by Lark (2009) in a case study where several covariates are available for modelling the non-stationarity of the mean and variance. The best stationary variance model is compared with the best non-stationary variance model, using evaluation criteria that measure both the quality of the predictions as well as the quality of the estimated prediction uncertainty.

2. Statistical methodology

2.1. Model definition

A soil variable of interest Z at any location $\mathbf{s}$ in the study area $\mathscr{A}$ is modelled by:

$$Z(\mathbf{s}) = m(\mathbf{s}) + \sigma(\mathbf{s})\varepsilon(\mathbf{s}) \tag{1}$$

where $m(\mathbf{s})$ is the mean at location $\mathbf{s}$, $\sigma(\mathbf{s})$ the standard deviation at location $\mathbf{s}$ and ε a stationary, spatially correlated Gaussian random field with zero mean and unit variance. The mean m and standard deviation σ are deterministic functions that are modelled as linear combinations of covariates, unconditional on the observations:

$$Z(\mathbf{s}) = \sum_{k=0}^{K} \beta_k w_k(\mathbf{s}) + \sum_{l=0}^{L} \kappa_l g_l(\mathbf{s})\varepsilon(\mathbf{s}) \tag{2}$$

where the β_k and κ_l are regression coefficients (the latter are used for modelling the standard deviation), and the w_k and g_l spatially distributed covariates. We take $w_0(\mathbf{s}) = g_0(\mathbf{s}) = 1$ for all $\mathbf{s}$, so that β_0 and κ_0 are space-invariant constant contributions to the mean and standard deviation, respectively.

Let Z be measured at n locations $\mathbf{s}_i (i = 1, \ldots, n; \mathbf{s}_i \in \mathscr{A})$. The measurements $z(\mathbf{s}_i)$ are treated as realizations of the Gaussian field Z and prediction is done for Z at a new, unobserved location $\mathbf{s}_0$. Stacking the $z(\mathbf{s}_i)$ in a (column) vector $\mathbf{z}$ and changing to matrix notation yields:

$$\mathbf{z} = \mathbf{W}\boldsymbol{\beta} + \mathbf{H}\boldsymbol{\varepsilon} \tag{3}$$

where $\mathbf{W}$ is the $n \times (K + 1)$ design matrix of covariates for the mean at the observation locations, $\boldsymbol{\beta}$ is the $(K + 1)$ vector of regression coefficients for the mean and $\boldsymbol{\varepsilon}$ is the n-vector of (standardized) residuals, which has variance-covariance matrix $\mathbf{R}$. $\mathbf{H}$ is an $n \times n$ diagonal matrix defined by:

$$\mathbf{H} = diag\{\mathbf{G}\boldsymbol{\kappa}\} \tag{4}$$

where $\mathbf{G}$ is the $n \times (L + 1)$ matrix of standard deviation covariates at observation locations and $\boldsymbol{\kappa}$ is an $(L + 1)$ vector of standard deviation regression coefficients. Note that while ε has variance-covariance matrix $\mathbf{R}$, the stochastic component $\mathbf{H}\boldsymbol{\varepsilon}$ of Eq. (3) has variance-covariance matrix $\mathbf{C} = \mathbf{HRH}'$. The parameters of the model defined by Eq. (3) are $\boldsymbol{\beta}$, $\boldsymbol{\kappa}$ and the parameters of a model for the spatial autocorrelation of the standardized residual. In this work we will parametrize the spatial autocorrelation by an isotropic exponential correlogram $r(h) = r_0\left\{\exp\left(-\frac{h}{a}\right)\right\}$ (where $h > 0$ is the Euclidean distance between two locations, by definition $r(0) = 1$), thus introducing two more parameters, namely r_0 and a. Parameter r_0 equals one minus the nugget-to-sill ratio, while a refers to the spatial correlation length (or range, $3a$ being the effective range). Note that the stationary variance model is a special case of the non-stationary variance model. It is obtained by setting parameters κ_l, $l = 1 \ldots L$ to zero, so that $\sigma(s) = \kappa_0$ for all $\mathbf{s}$.

2.2. Parameter estimation and model selection

2.2.1. Parameter estimation

In estimation the parameters are subdivided in two subsets, the regression coefficients $\boldsymbol{\beta}$ for the mean, and all parameters of the stochastic part of the model, $\Phi = [\boldsymbol{\kappa}, r_0, a]$. For a stationary variance model the second subset reduces to $\Phi = [\kappa_0, r_0, a]$. The standard maximum likelihood estimates of Φ depend non-linearly on the regression coefficients for the mean $\boldsymbol{\beta}$, which introduces a bias in the estimates of Φ if both parameter subsets are estimated jointly (Lark and Webster, 2006). This problem can be avoided by restricted (or residual) maximum likelihood (REML) parameter estimation. REML first estimates Φ and next $\boldsymbol{\beta}$. Similar to standard maximum likelihood estimation, REML aims to find the vector Φ for which the observed data yield the highest probability density (i.e., likelihood, if treated as a function of the parameters instead of the data). The problem is that the likelihood of Φ depends on the regression coefficients for the mean, which are unknown and must also be estimated. Patterson and Thompson (1971) solved this problem by detrending the data by multiplying the data vector by a projection matrix. The new variable is a function of the original variable but independent of the regression coefficients for the mean. The associated restricted log-likelihood function is given by (Webster and Oliver, 2007):

$$L_r(\Phi|\mathbf{z}) = constant - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}\log|\mathbf{W}'\mathbf{C}^{-1}\mathbf{W}| - \frac{1}{2}\mathbf{z}'\mathbf{P}'\mathbf{C}^{-1}(\mathbf{I} - \mathbf{Q})\mathbf{z} \tag{5}$$

where $\mathbf{I}$ is an identity matrix and $\mathbf{P}$ and $\mathbf{Q}$ are defined as:

$$\mathbf{P} = \mathbf{I} - \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}' \tag{6}$$

$$\mathbf{Q} = \mathbf{W}(\mathbf{W}'\mathbf{C}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{C}^{-1} \tag{7}$$

After estimating Φ by maximising the restricted log-likelihood given in Eq. (5) above, the regression coefficients $\boldsymbol{\beta}$ for the mean can be estimated by Generalized Least Squares (GLS):

$$\boldsymbol{\beta} = (\mathbf{W}'\mathbf{C}^{-1}\mathbf{W})^{-1}\mathbf{W}'\mathbf{C}^{-1}\mathbf{z} \tag{8}$$

Here, matrix $\mathbf{C}$ is computed from the optimized values for Φ. Note that the regression coefficients κ_l in Eq. (2) can be positive or negative, as long as the covariance matrix $\mathbf{C}$ is not singular.

2.2.2. Model selection

Two subsets of covariates must be chosen, one for the mean and one for the standard deviation. Suppose we have in total K candidate covariates for modelling the mean. For a subset of covariates of size k, there are $\binom{K}{k}$ possible combinations. Since the size is not fixed, we have $\sum_{k=0}^{K}\binom{K}{k}$ possible models in total for the stationary variance model.