Geoderma 262 (2016) 174–186

Contents lists available at ScienceDirect

Geoderma

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

A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals

T.G. Orton [a,b,*], M.J. Pringle [b], T.F.A. Bishop [a]

[a] Faculty of Agriculture and Environment, The University of Sydney, 1 Central Avenue, Australia Technology Park, Eveleigh, NSW 2015, Australia
[b] EcoSciences Precinct, Department of Science, Information Technology and Innovation, GPO Box 5078, Brisbane, QLD 4001, Australia

ARTICLE INFO

Article history:
Received 16 April 2015
Received in revised form 31 July 2015
Accepted 10 August 2015
Available online 29 August 2015

Keywords:
Geostatistics
Spatial variability
Soil profile
Depth function
Area-to-point kriging

ABSTRACT

Datasets for modelling and mapping soil properties often consist of samples from many spatial locations, collected from several different soil depth intervals. However, interest may lie in the spatial distribution of the property for a particular target depth interval, which may or may not correspond to the sampled intervals. It is the task of the data analyst to put the data together in such a way that useful and reliable conclusions can be drawn for the soil depths of practical interest. Previous studies to tackle this problem include multi-stage approaches and point-data-based 3-dimensional geostatistical approaches. One disadvantage of a multi-stage approach — for example, first fitting splines to the data for sampled profiles, then imputing new data for the target interval, before considering a spatial analysis with the imputed data — is that the imputation generally ignores any uncertainty in the imputed data, which might give misleading conclusions. Point geostatistical methods, on the other hand, assume that the data represent the value of the target variable at a specific point in the profile, rather than its average over a sampling interval; this too could give misleading estimates. In this work, we present a statistical method that properly deals with the sample support of soil profile data so that all data can be considered in a single geostatistical analysis. The approach is based on the 'area-to-point' kriging framework, which can be used to represent the uncertainty from data that are averages over non-negligible sample supports (in our case, the different sampled depth intervals). We combine a covariance model for the increment-averaged data in the vertical domain with another model for the horizontal variation. This enables us to (i). process all data in a single analysis, and (ii). calculate predictions for any target depth and support based on the same statistical model. We test the approach on data from the Murray–Darling basin in eastern Australia, where interest lies in mapping various soil properties that could have an effect on water salinity of the nearby Muttama Creek: we illustrate the methodology for predicting clay content. Finally we discuss a number of possible extensions of the methodology to broaden its applicability, which should provide the basis of further studies.

1. Introduction

Soil properties vary significantly both across the landscape and through the soil profile, and interest lies in characterizing and mapping this variation to provide land users with useful information. Datasets often consist of samples from many spatial locations, at several different depth intervals. Within a particular study, these depth intervals may be fixed (e.g. 0–10 cm, 10–20 cm, and 20–30 cm). Other studies may consider different fixed intervals, or sampling intervals that are defined according to soil horizons and therefore vary between locations within the study. It is then the task of the data analyst to draw useful and reliable conclusions for soil depths of practical interest. For example, the GlobalSoilMap project specifications (Arrouays et al., 2014) dictate that soil properties should be mapped for depth intervals of 0–5 cm, 5–15 cm, 15–30 cm, 30–60 cm, 60–100 cm and 100–200 cm.

Various approaches have been adopted previously to perform this task. A common approach is to fit splines to the profile data for each site (Bishop et al., 1999), use the spline to impute 'data' for the soil property over the depth interval of interest, and then proceed with the analysis as if the value were known without error (e.g. Malone et al., 2009, 2011a; Adhikari et al., 2013; Orton et al., 2014; Bishop et al., 2015). We refer to this as a 'spline-then-krige' (STK) approach. This process does not account for the uncertainty in the values inferred from the spline, and could yield misleading conclusions.

Another possible approach to the problem is 3-dimensional (3-D) geostatistics. However, this has been applied as if the data collected from soil depth intervals were concentrated at a single point (e.g. Hengl et al., 2014), at either one of the bounds of the sampling interval, or at the interval's mid-point. This approach also fails to properly represent the support on which the data were originally collected (over an interval, rather than from a point), and could again yield misleading conclusions. Breidt et al. (2007) recognized the dangers of using mid-point assignment to represent increment-averaged data, and proposed

* Corresponding author.
E-mail address: Thomas.Orton@dsiti.qld.gov.au (T.G. Orton).

http://dx.doi.org/10.1016/j.geoderma.2015.08.013
0016-7061/

a mixed-model approach for estimating depth functions, whilst properly accounting for the interval support of the data; their focus was on the estimation of depth profiles, whereas our focus is more on the use of such data for modelling and mapping using spatial datasets of soil horizon data. Other 3-D approaches (e.g. Poggio and Gimona, 2014; Veronesi et al., 2012) generally suffer the same drawback; all data are assumed to have identical vertical support, which ignores their different uncertainties.

In a geostatistical framework, the sample support of data that are averages of an attribute over non-negligible areal units can be dealt with by area-to-point kriging (ATP kriging; Kyriakidis, 2004). This method allows the sampling units and prediction supports to all have different sizes and shapes. It has been applied in several case studies in recent years to analyse areal-averaged data (Kyriakidis and Yoo, 2005; Kerry et al., 2012; Schirrmann et al., 2012; Truong et al., 2014). Although usually carried out to account for the horizontal support (i.e. the data are areal averages), there is no reason that the same methodology cannot be carried out to deal with the vertical support of soil profile data (i.e. for data that are measurements of the average value of a soil property over depth intervals). This was noted in Heuvelink (2014), although we are unaware of any studies that have implemented such an approach.

In this work, we combine the ATP approach for the vertical distribution with standard kriging approaches for the horizontal distribution. Thus, a statistical model for the complete dataset (all spatial locations and all depth intervals) is defined, with the support of each datum (a combination of spatial location and depth interval) properly represented. We refer to this model for increment-averaged data, and the predictions built on the model, as increment-averaged kriging (IAK). We propose that this all-in-one model should provide a better assessment of prediction uncertainty compared with a two-stage approach, or an approach that represents interval data by their mid-points (although comparison of the different methodologies is not undertaken in the current study).

We consider the methodology in the framework of a linear mixed model (LMM; Lark et al., 2006). Thus, part of the variation of the target variable can be explained by a collection of explanatory variables, with the remainder being modelled as spatially dependent (i.e. data close to each other in horizontal space and at similar depths are more likely to be similar than data far apart in space and at different depths). We allow interactions between depth and the spatial explanatory variables, so that different relationships can be modelled at different depths in the profile. We also allow the variance parameters of residuals to depend on depth, which provides a mechanism to represent different uncertainties at different depths in the profile.

Usually in ATP-kriging studies, the average covariances must be calculated numerically (by a discretization approach), due to the complex nature of the areal data units in 2-D space. However, for our increment-averaged data, the average covariances can be computed analytically. We derive an expression for the covariance of the increment-averaged data, based on an exponential model for the point covariances. This significantly reduces the computational load of maximum likelihood methods compared with numerical procedures. Nonetheless, for large datasets (when the total number of data is more than a few thousand), likelihood approximation techniques may have to be used (e.g. Stein et al., 2004; Eidsvik et al., 2014); we do not consider these here though.

We test the proposed IAK approach on data from the Murray–Darling basin in eastern Australia, where interest lies in mapping soil properties that could have an effect on water salinity of the nearby Muttama Creek. Soil cores (to a depth of 1 m) were collected from 55 spatial locations over the Muttama catchment, and each core was divided into horizons, giving a total of 192 samples. We use this case study to illustrate the IAK approach, mapping clay content and its attendant uncertainty based on the data from these samples.

2. Theory

Throughout the following, we will assume that the horizontal support of the data and of the prediction is point support. The method can be extended to deal with data that are both areal- and depth-wise averages, if this were to be required in another study. We begin our presentation of the methodology with a simple stationary model for the point covariances. We then extend this model to a more realistic one, allowing variances to depend on depth, before describing how this relates to the average covariances required to model the variation of increment-averaged data.

2.1. IAK model: initial stationary model for point covariances

We begin our development towards a statistical model for the analysis of depth interval-averaged data by considering a 3-D model for point data (i.e. with depths, d, taken to be fixed points):

$$y(\mathbf{x}, d) = \mu(\mathbf{x}, d) + \varepsilon(\mathbf{x}, d) \tag{1}$$

where $\mathbf{x}$ are the horizontal coordinates. We use a linear model of some known covariates to give the trend function, $\mu(\mathbf{x}, d)$, which can be written as:

$$\mu(\mathbf{x}, d) = \mathbf{X}(\mathbf{x}, d)\boldsymbol{\beta} \tag{2}$$

where $\mathbf{X}(\mathbf{x}, d)$ contains the known values of the covariates and $\boldsymbol{\beta}$ is the vector of associated parameters (to be estimated). This is known as the fixed-effect function, and $\mathbf{X}(\mathbf{x}, d)$ constitutes a row of the fixed-effect design matrix. We assume that the residuals, $\varepsilon(\mathbf{x}, d)$, follow a multivariate normal distribution with mean zero and covariances depending only on the horizontal and vertical separation distances (this assumption will be relaxed in Section 2.2). As a first approach, we assume a separable (product) covariance model (De Iaco et al., 2011):

$$\mathrm{Cov}[Y(\mathbf{x}, d), Y(\mathbf{x}', d')] = \sigma^2 \ \varphi_{\mathbf{x}}(h_{\mathbf{x}}; \boldsymbol{\vartheta}_{\mathbf{x}})\varphi_d(h_d; \boldsymbol{\vartheta}_d) \tag{3}$$

where $\varphi_{\mathbf{x}}(h_{\mathbf{x}};\ \boldsymbol{\vartheta}_{\mathbf{x}})$ is a correlation function for any pair of observations separated by distance $h_{\mathbf{x}} = |\mathbf{x}' - \mathbf{x}|$, and $\varphi_d(h_d;\ \boldsymbol{\vartheta}_d)$ is another correlation function of the vertical separation distances, $h_d = |d' - d|$. The parameter, σ^2, is the variance. This simple model assumes that the covariances can be written as the product of a function that depends only on the horizontal separation distances, and one that depends only on the vertical separation distances. Although this can be restrictive, it provides a useful starting point, and we suggest possible alternatives in the discussion.

In the product covariance model, we can choose any permissible correlation functions (see e.g. Webster and Oliver, 2001) for $\varphi_{\mathbf{x}}(h_{\mathbf{x}};\ \boldsymbol{\vartheta}_{\mathbf{x}})$ and $\varphi_d(h_d;\ \boldsymbol{\vartheta}_d)$. However, as Truong et al. (2014) point out, there is no information in areal-averaged data (or in our case, depth interval-averaged data) to define a nugget effect. We therefore assume that the depth-wise correlation function, $\varphi_d(h_d;\ \boldsymbol{\vartheta}_d)$, has zero nugget, and model it with a single spatial autocorrelation structure. We can still include a nugget effect for the horizontal variation though, and we write $\varphi_{\mathbf{x}}(h_{\mathbf{x}};\ \boldsymbol{\vartheta}_{\mathbf{x}})$ as the sum of a nugget component and N_m spatial autocorrelation structures:

$$\varphi_{\mathbf{x}}(h_{\mathbf{x}};\ \boldsymbol{\vartheta}_{\mathbf{x}}) = s_0\, \varphi_{\mathbf{x};0}(h_{\mathbf{x}}) + \sum_{i=1}^{N_m} s_i\, \varphi_{\mathbf{x};i}(h_{\mathbf{x}};\ \boldsymbol{\vartheta}_{\mathbf{x};i}) \tag{4}$$

where: $\varphi_{\mathbf{x};0}(h_{\mathbf{x}}) = \begin{cases} 1 & \text{if } h_{\mathbf{x}} = 0 \\ 0 & \text{otherwise} \end{cases}$ is the nugget correlation function;

$\varphi_{\mathbf{x};i}(h_{\mathbf{x}};\ \boldsymbol{\vartheta}_{\mathbf{x};i}), i = 1, \ldots, N_m$, are N_m spatial correlation functions, with parameter vectors $\boldsymbol{\vartheta}_{\mathbf{x};i}$; parameters $s_i, i = 1, \ldots, N_m$ are the proportions of variance associated with each of the N_m spatial correlation functions;