Geoderma 173-174 (2012) 241–248

Contents lists available at SciVerse ScienceDirect

Geoderma

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

A hybrid design-based and model-based sampling approach to estimate the temporal trend of spatial means

D.J. Brus *, J.J. de Gruijter

Alterra, Wageningen University and Research Centre, PO Box 32, 6700 AA Wageningen, The Netherlands

ARTICLE INFO

Article history:
Received 7 April 2011
Received in revised form 12 October 2011
Accepted 10 December 2011
Available online 4 February 2012

Keywords:
Soil monitoring
Space–time sampling
Rotating panel
Model error
Linear mixed model
REML

ABSTRACT

This paper launches a hybrid sampling approach, entailing a design-based approach in space followed by a model-based approach in time, for estimating temporal trends of spatial means or totals. The underlying space–time process that generated the soil data is only partly described, viz. by a linear mixed model for the temporal variation of the spatial means. The model contains error terms for model inadequacy (model or process error) and for the sampling error in the estimated spatial means. The linear trend is estimated by Generalized Least Squares. The covariance matrix is obtained by adding the matrix with design-based estimates of the sampling variances and covariances and the covariance matrix of the model errors. The model parameters needed for the latter matrix are estimated by REML. The error variance of the estimated regression coefficients can be decomposed into the model variance of the errorless regression coefficients and the model expectation of the conditional sampling variance. In a case study on forest soil eutrophication, inclusion of the model error led to a considerable increase of the error variance for most variables. In the topsoil the contribution of the process error to the standard error of the estimated trend was much larger than that of the sampling error. For pH there was no contribution of the model error. Important advantages of the presented approach over the fully model-based approach are its simplicity and robustness to model assumptions.

1. Introduction

A major decision in designing sampling schemes for soil monitoring is the choice between a design-based and a model-based sampling strategy (Brus and de Gruijter, 1993, 1997; de Gruijter and ter Braak, 1990; Papritz and Webster, 1995). In the design-based approach sampling units are selected by probability sampling, and the inference is based on the sampling design. In a model-based strategy sampling units need not be selected by probability sampling, and are generally selected purposively, for instance such that they are well spread out in geographic space (van Groenigen et al., 1999; Walvoort et al., 2010) and/or in feature (predictor) space (Brus and Heuvelink, 2007). The statistical inference is based on a stochastic model of variation of the property of interest in space and/or time.

When sampling in space and time, in principle both sampling locations and sampling times can be selected by probability sampling. In this case, a model of variation is not needed for statistical inference, but the inference can be entirely based on the spatial and temporal sampling designs. This fully design-based approach can be advantageous in compliance monitoring of the space–time mean or space–time total, e.g. the total annual CO_2 emission in a region. In compliance monitoring the aim is to decide (by statistical testing) whether the sampling universe satisfies regulatory conditions. In the fully design-based approach no model of variation is used, which enhances the validity of the result. In compliance monitoring validity of the result is of special importance in order to avoid a hard to settle debate on whether the status of the soil complies with the (legislative) standard or not. See Brus and Knotters (2008) for an application on compliance monitoring of water quality.

As opposed to the fully design-based approach, in the fully model-based approach the inference is based on a stochastic model of the variation in space and time, and consequently neither sampling times nor sampling locations need to be selected by probability sampling. ter Braak et al. (2008) derived the Best Linear Unbiased Predictor (BLUP) for the linear temporal trend of the spatial mean and its variance under a universal kriging model (linear mixed model) in which the variance of the residuals is modeled by a space–time variogram with geometric anisotropy. This universal kriging predictor can be used, for instance, to estimate the temporal trend of the spatial means of soil properties such as carbon stocks and pH from legacy data that usually are not collected from probability samples.

In this paper we introduce and demonstrate a hybrid, design-based and model-based approach for sampling in space and time. In this hybrid approach sampling locations are selected by probability sampling, whereas times are not. We will show that, contrary to the fully model-based approach in which the variation in both space

* Corresponding author. Tel.: + 31 317 486250; fax: + 31 317 419000.
E-mail addresses: dick.brus@wur.nl (D.J. Brus), jaap.degruijter@wur.nl (J.J. de Gruijter).

0016-7061/$ – see front matter
doi:10.1016/j.geoderma.2011.12.010

and time is described by a model, in the hybrid approach the stochastic space–time process is only partly described, namely by a model of the temporal variation of the spatial means only. In quantifying the uncertainty about the target parameter, two stochastic processes are accounted for, the random selection of the sampling locations *and* the stochastic space–time process.

In this paper we focus on estimation of the temporal trend of the spatial mean defined as a model parameter (regression coefficient) under the hybrid approach. We will demonstrate the hybrid approach with a case study on acidification and eutrophication of forest soils. The results obtained with the hybrid approach will be compared with the results obtained with the design-based approach as reported by Brus and de Gruijter (2011). We will discuss the advantages of the hybrid sampling approach over the fully model-based approach. We will argue that if the monitoring data are yet to be collected and interest is in global target quantities such as the temporal trend of spatial means, then the hybrid sampling approach can be advantageous because a full space–time model need not be identified. Especially with sparse data the calibration of such a model can be challenging.

2. Theory

2.1. Time series model for spatial means

The hybrid sampling approach is based on the publication of Jones (1980) who developed a general framework for estimating the population means at multiple sampling times under the time series approach, see also Binder and Hidiroglou (1988), p. 201 for an excellent review. In this approach the population means are modeled as random variables, not as fixed population parameters as in the design-based approach. Besides model errors, sampling errors in the estimated population means are accounted for in the statistical inference, obtained by design-based inference from probability samples. We therefore refer to this approach as the hybrid, design-based and model-based sampling approach.

In this hybrid approach a model is postulated for the temporal variation of the spatial mean, total or fraction. As applications to these parameters are completely similar, we confine our description of the approach further to the mean. The spatial mean of the target variable at time t_j, $\bar{Y}(t_j)$, is defined as:

$$\bar{Y}\left(t_j\right) = \frac{1}{\|\mathcal{A}\|}\int_{s\epsilon\mathcal{A}} Y\left(\mathbf{s}, t_j\right)\mathrm{d}\mathbf{s} \tag{1}$$

In this study we adopt a linear mixed model for the space–time process ξ:

$$\bar{Y}\left(t_j\right) = \sum_{u=1}^{q}\beta_u d_u\left(t_j\right) + \eta\left(t_j\right) \tag{2}$$

with $d_u(t_j)$ the u^{th} predictor at time t_j ($j = 1 \cdots r$), β_u the regression coefficient for this predictor, and $\eta(t_j)$ the model residual of the spatial mean at time t_j, also referred to as the model error or the process error. The predictors can be a constant with value 1 (for the intercept), the time t (see hereafter), or an explanatory variable related to the variable of interest.

In practice the spatial means are unknown, and in the hybrid approach these means are estimated from a probability sample, for instance by the Horvitz–Thompson estimator:

$$\hat{\bar{Y}}\left(t_j\right) = \frac{1}{\|\mathcal{A}\|}\sum_{l=1}^{n(t_j)}\frac{Y_l\left(t_j\right)}{\pi_l} \tag{3}$$

with $n(t_j)$ the number of sampling locations at time t_j, and π_l the inclusion density of sampling location l. We consider the situation where we can have several 'elementary' estimates of the spatial mean at a given time t_j. An elementary estimate is an estimate from one panel, i.e. from one set of locations observed at the same set of times (Brus and de Gruijter, 2011).

The sampling introduces an additional error component in the model for the i^{th} elementary estimate of the spatial mean at time t_j, $\hat{\bar{Y}}_i(t_j)$:

$$\hat{\bar{Y}}_i\left(t_j\right) = \sum_{u=1}^{q}\beta_u d_u\left(t_j\right) + \eta\left(t_j\right) + \epsilon_i\left(t_j\right) \tag{4}$$

with $\boldsymbol{\epsilon}_i(t_j)$ the sampling error of the i^{th} elementary estimate of the spatial mean at time t_j. If we take in Eq. (4) $d_1(t_j) = 1$ and $d_2(t_j) = t_j$ for $j = 1 \cdots r$, the linear mixed model becomes:

$$\hat{\bar{Y}}_i\left(t_j\right) = \beta_1 + \beta_2 \cdot t_j + \eta\left(t_j\right) + \varepsilon_i\left(t_j\right) \tag{5}$$

where β_2 is the model parameter describing the linear temporal trend of the spatial mean. Note that if we take $x_1(t_j) = 1$ and $x_2(t_j) = I$, a 0/1 indicator indicating whether a sampling round takes place before or after some event, the model describes a step-trend, which might be more relevant in effect-monitoring. In matrix notation Eq. (5) becomes

$$\hat{\bar{\mathbf{Y}}} = \mathbf{D}\boldsymbol{\beta} + \mathbf{X}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{6}$$

with $\mathbf{D}$ the $L \times 2$ matrix with 1's in first column and the sampling times $t_1 \cdots t_r$ in the second column (L is the total number of elementary estimates: $L = \sum_j l_{t_j}$), and $\mathbf{X}$ the $L \times r$ matrix with 0's and 1's selecting the appropriate elements from $\boldsymbol{\eta}$. We extend the model with the following probability model for the errors $\boldsymbol{\eta}$ and $\boldsymbol{\epsilon}$:

$$\begin{bmatrix}\boldsymbol{\eta}\\ \boldsymbol{\epsilon}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\mathbf{C}_\xi & 0\\ 0 & \mathbf{C}_p\end{bmatrix}\right)$$

The model errors $\boldsymbol{\eta}$ have zero mean and an $r \times r$ covariance matrix $\mathbf{C}_\xi$. The sampling errors have zero mean and an $L \times L$ covariance matrix $\mathbf{C}_p$. Subscript p refers to the sampling design used to select the locations. The covariances of the model error $\boldsymbol{\eta}$ and sampling error $\boldsymbol{\epsilon}$ equal zero, as they originate from independent stochastic processes. The overall covariance matrix of the estimated spatial means therefore equals

$$\mathbf{C}_{\xi p} = \mathbf{X}\mathbf{C}_\xi\mathbf{X}' + \mathbf{C}_p \tag{7}$$

2.2. Estimation of regression coefficients with known covariance matrix $\mathbf{C}_{\xi p}$

With known covariance matrix $\mathbf{C}_{\xi p}$,the regression coefficients can be estimated by Generalized Least Squares (GLS):

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\mathbf{D}\right)^{-1}\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\hat{\bar{\mathbf{Y}}} \tag{8}$$

This GLS estimator is equal to the maximum likelihood estimator given the matrix $\mathbf{C}_{\xi p}$ (Diggle and Ribeiro, 2007).

Variance of estimated regression coefficients. The covariance matrix of the estimated regression coefficients can be obtained by

$$\mathrm{Var}\left(\hat{\boldsymbol{\beta}}\right) = \left(\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\mathbf{D}\right)^{-1} \tag{9}$$

The variance of an estimated regression coefficient can be decomposed as follows:

$$\mathrm{Var}\left(\hat{\beta}\right) = \mathrm{Var}_\xi\left\{\mathrm{E}_p\left(\hat{\beta}\right)|\xi_0\right\} + \mathrm{E}_\xi\left\{\mathrm{Var}_p\left(\hat{\beta}\right)|\xi_0\right\} \tag{10}$$