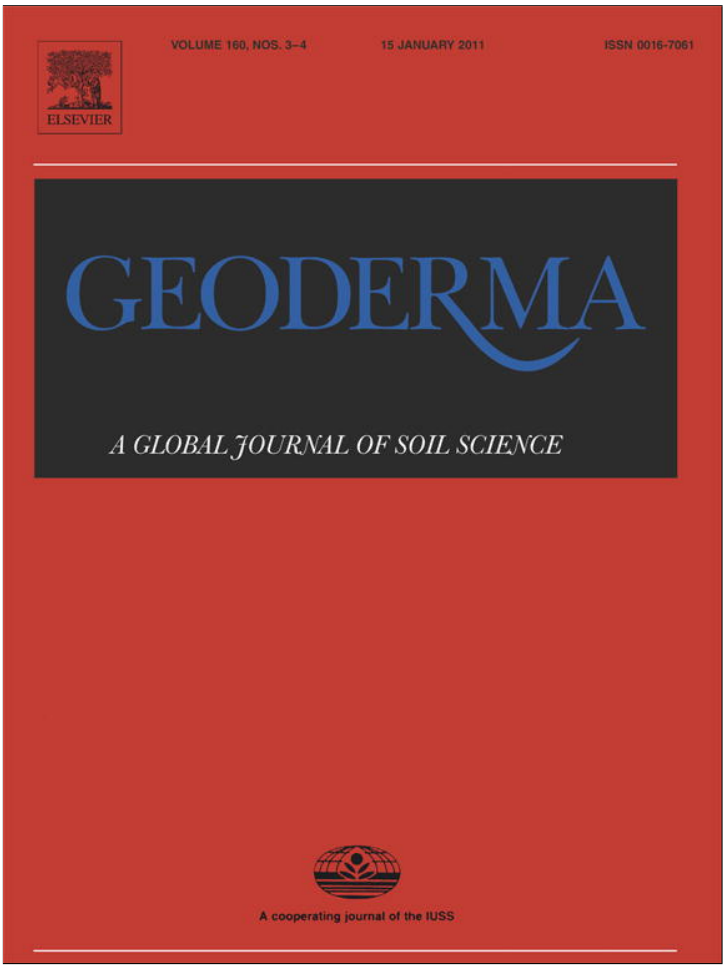

(This is a sample cover image for this issue. The actual cover is not yet available at this time.)

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 xxxx

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 © 2011 Elsevier B.V. All rights reserved.
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:

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

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):

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

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:

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

with $\mathrm{Var}_\xi\left\{\mathrm{E}_p\left(\hat{\boldsymbol{\beta}}\right)|\xi_0\right\}$ the model variance of the conditional p-expectation of the regression coefficients (conditioned on a realization of the space–time process), and $\mathrm{E}_\xi\left\{\mathrm{Var}_p\left(\hat{\boldsymbol{\beta}}\right)|\xi_0\right\}$ the model expectation of the conditional sampling variance of the estimated regression coefficients. With errorless observations at the sampling locations and a consistent estimator of the spatial means, the model variance of the conditional p-expectation of the regression coefficients equals

$$\mathrm{Var}_\xi\left\{\mathrm{E}_p\left(\hat{\boldsymbol{\beta}}\right)|\xi_0\right\} = \left(\mathbf{D}'_r\mathbf{C}_\xi^{-1}\mathbf{D}_r\right)^{-1}, \tag{11}$$

with $\mathbf{D}_r$ the reduced $(r\times 2)$ design matrix with ones in first column and $t_1 \cdots t_r$ in the second column. This is the model variance of the errorless regression coefficients, i.e. the regression coefficients that would have been obtained with exhaustive spatial sampling at all selected time points. The model expectation of the conditional sampling variance of the estimated regression coefficients equals

$$\mathrm{E}_\xi\left\{\mathrm{Var}_p\left(\hat{\boldsymbol{\beta}}\right)|\xi_0\right\} = \mathrm{E}_\xi\left\{\left(\mathbf{D}'\mathbf{C}_p^{-1}\mathbf{D}\right)^{-1}|\xi_0\right\} \tag{12}$$

This decomposition shows that the model variance of the regression coefficients (Eq. (11)) is fully determined by the model variance parameters and the sampling times. The effect of the space–time sampling design on the variance of the regression coefficients goes via the second term in Eq. (10), the model expectation of the conditional sampling variance of the estimated regression coefficients (Eq. (12)).

2.3. Estimation of regression coefficients with unknown covariance matrix $C_{\xi p}$

In practice the matrix $\mathbf{C}_{\xi p}$ is unknown and must be estimated from the sample data. Covariance matrix $\mathbf{C}_{\xi p}$ can be determined by estimating the sampling covariance matrix $\mathbf{C}_p$ and the model covariance matrix $\mathbf{C}_\xi$, and then adding these two matrices (Eq. (7)). The estimated covariance matrix is then plugged into the GLS estimator of the regression coefficients (Eq. (8)).

Estimation of sampling covariance matrix $\mathbf{C}_p$ .The sampling variances and covariances (elements of matrix $\mathbf{C}_p$) can be estimated by the well-known design-based variance estimators (de Gruijter et al., 2006). For instance, with simple random sampling (with replacement) the sampling variance can be estimated by:

$$\mathrm{Var}_j = \frac{\hat{\sigma}_j^2}{m} \tag{13}$$

with m the number of sampling locations in the panel on which the elementary estimate is based, and

$$\hat{\sigma}_j^2 = \frac{1}{n-1}\sum_{u=1}^{n}\left(y_{ju}-\hat{\bar{y}}_j\right)^2 \tag{14}$$

with $\hat{\bar{y}}_j$ the estimated mean (sample mean) at time t_j. Note that all sampling locations of time t_j can be used to estimate the spatial variance. The sampling covariance of elementary estimates from independently selected panels equals 0. The sampling covariance of elementary estimates from the same panel equals

$$\mathrm{Cov}_{jk} = \frac{\hat{\sigma}_{jk}^2}{m} \tag{15}$$

with

$$\hat{\sigma}_{jk}^2 = \frac{1}{m-1}\sum_{u=1}^{m}\left(y_{ju}-\hat{\bar{y}}_j^{(m)}\right)\left(y_{ku}-\hat{\bar{y}}_k^{(m)}\right) \tag{16}$$

with $\hat{\bar{y}}_j^{(m)}$ the sample mean in the panel.

Estimation of model covariance matrix $\mathbf{C}_\xi$. We now proceed with estimation of the variances and covariances of the model errors (elements in matrix $\mathbf{C}_\xi$). Given an authorized covariance function for the spatial means $C(t)=f(t,\boldsymbol{\theta})$, the parameters $\boldsymbol{\theta}$ of this function can be estimated by residual maximum likelihood (Lark et al., 2006). The residual negative log likelihood equals (Diggle and Ribeiro, 2007):

$$L_R\left(\boldsymbol{\theta},\boldsymbol{\beta}|\hat{\mathbf{Y}}\right) = \frac{1}{2}\left\{n\log(2\pi)+\log|\mathbf{C}_{\xi p}|+\log|\mathbf{D}'\mathbf{C}_{\xi p}^{-1}\mathbf{D}|+\left(\hat{\mathbf{Y}}-\mathbf{D}\hat{\boldsymbol{\beta}}\right)'\mathbf{C}_{\xi p}^{-1}\left(\hat{\mathbf{Y}}-\mathbf{D}\hat{\boldsymbol{\beta}}\right)\right\} \tag{17}$$

with $\hat{\boldsymbol{\beta}}$ the maximum likelihood estimator of $\boldsymbol{\beta}$ given the parameters $\boldsymbol{\theta}$, which is given by Eq. (8). By minimizing Eq. (17) we obtain REML estimates of $\boldsymbol{\theta}$. These estimates $\hat{\boldsymbol{\theta}}_{\mathrm{REML}}$ can then be used to estimate the regression coefficients $\boldsymbol{\beta}$ (Eq. (8)).

3. Case study

We demonstrate the hybrid sampling approach with a case study on forest soil eutrophication and acidification. The data of this study were formerly used to estimate linear temporal trend of spatial means of soil properties by a design-based approach. Here we describe the case study briefly, for more details we refer the reader to Brus and de Gruijter (2011). An existing network was subsampled by simple random sampling without replacement. The sample size n was 20. Sampling was repeated four times ($r=4$), with an interval of approximately 1 year (2004-2007) according to a rotational design. The matching proportion was 0.5, which means that 10 randomly selected locations from a given sampling time were revisited at the next sampling time. In the fourth sampling time the 10 new locations of the third sampling time were revisited, plus the 10 locations of the first sampling time that were *not* revisited in the second sampling time (Fig. 1). The total number of panels thus is four, and the total number of observed plots equals 40.

For each time we have two elementary estimates of the spatial mean, so the total number of elementary estimates equals eight. Let $\hat{\bar{y}}_{pj}$ denote the sample mean at time t_j, with $j=1\cdots 4$ in panel p, with $p\in(a,b,c,d)$. If the eight elementary estimates are ordered as $(\hat{\bar{y}}_{a4},\hat{\bar{y}}_{a1},\hat{\bar{y}}_{b1},\hat{\bar{y}}_{b2},\hat{\bar{y}}_{c2},\hat{\bar{y}}_{c3},\hat{\bar{y}}_{d3},\hat{\bar{y}}_{d4})$, with $\hat{\bar{y}}_{a4}$ the elementary estimate at time t_4 from panel a, then the design matrix $\mathbf{X}$ equals

$$\begin{bmatrix}0&0&0&1\\1&0&0&0\\1&0&0&0\\0&1&0&0\\0&1&0&0\\0&0&1&0\\0&0&1&0\\0&0&0&1\end{bmatrix}$$

and the sampling covariance matrix $\mathbf{C}_p$ equals

$$\begin{bmatrix}\mathrm{Var}_{a4}&\mathrm{Cov}_{1,4}&0&0&0&0&0&0\\\mathrm{Cov}_{4,1}&\mathrm{Var}_{a1}&0&0&0&0&0&0\\0&0&\mathrm{Var}_{b1}&\mathrm{Cov}_{1,2}&0&0&0&0\\0&0&\mathrm{Cov}_{2,1}&\mathrm{Var}_{b2}&0&0&0&0\\0&0&0&0&\mathrm{Var}_{c2}&\mathrm{Cov}_{2,3}&0&0\\0&0&0&0&\mathrm{Cov}_{3,2}&\mathrm{Var}_{c3}&0&0\\0&0&0&0&0&0&\mathrm{Var}_{d3}&\mathrm{Cov}_{3,4}\\0&0&0&0&0&0&\mathrm{Cov}_{4,3}&\mathrm{Var}_{d4}\end{bmatrix}$$

The sample data were statistically analyzed as if they were selected by (one-stage) simple random sampling from an *infinite* population,

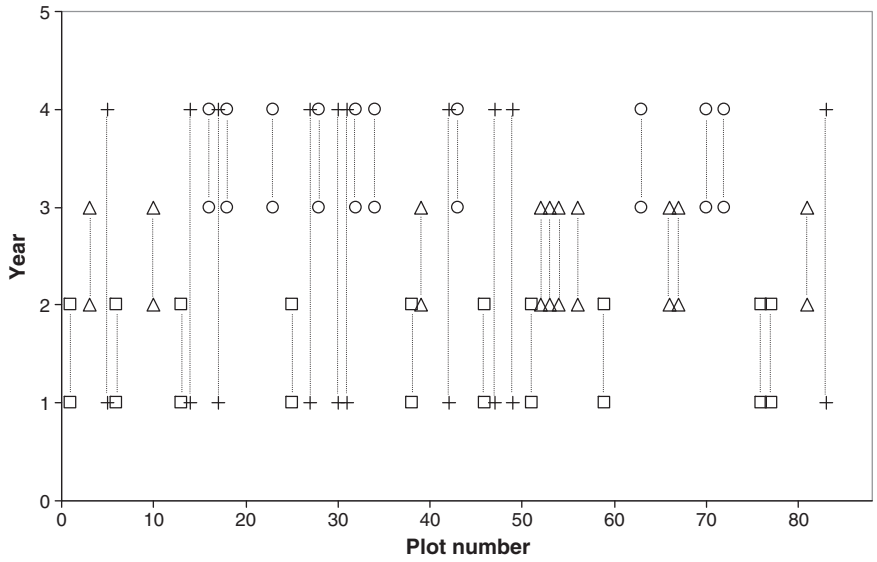

Fig. 1. Schematic representation of rotational sample as applied in forest soils of Utrechtse Heuvelrug. Revisits are indicated by vertical dotted lines.

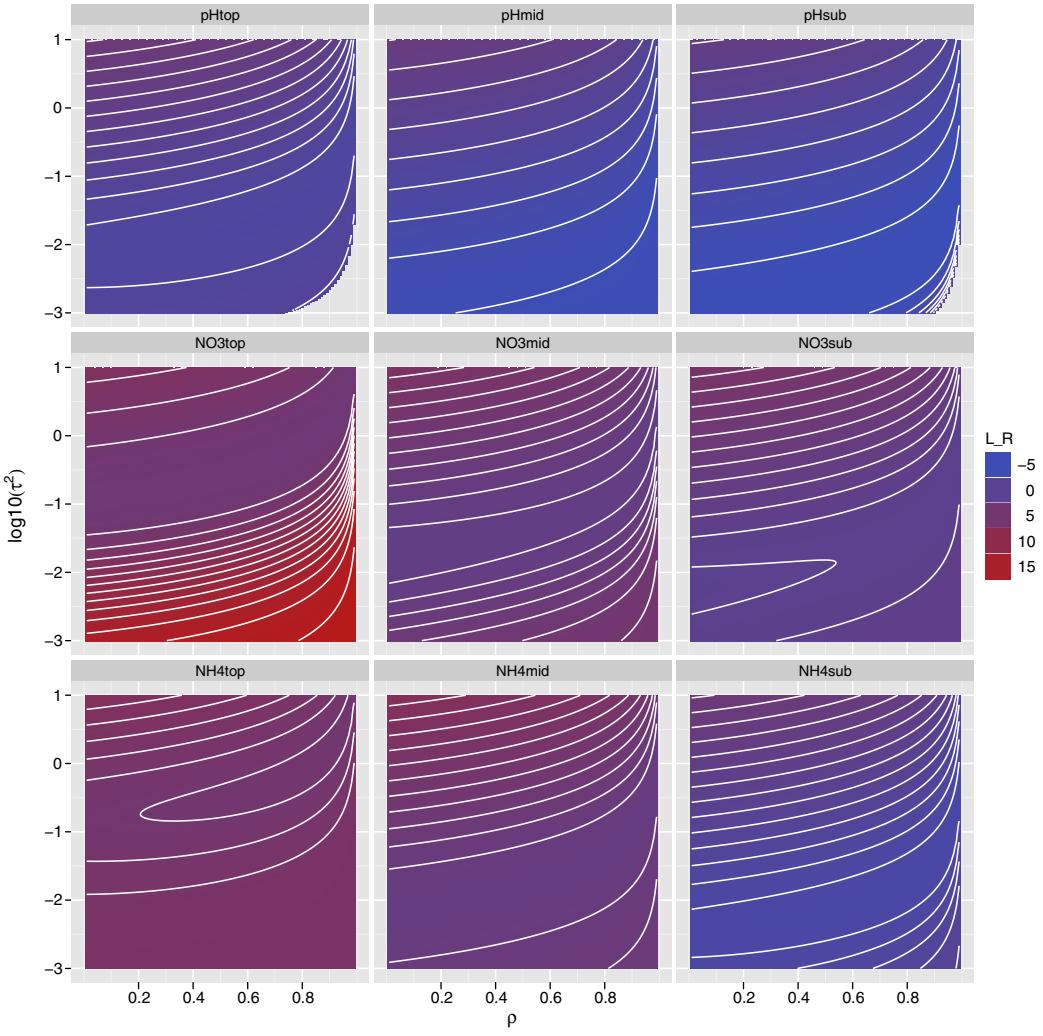

Fig. 2. Countour plots of the negative residual loglikelihood.

Table 1
REML estimates of the model variance (τ^2) and the autoregressive parameter (ρ).

Comp.	Depth	τ^2	ρ
pH	Top	0.851	0.99
pH	Mid	0.000	0.99
pH	Sub	0.110	0.99
NO_3	Top	7.079	0.98
NO_3	Mid	0.017	0.01
NO_3	Sub	0.005	0.01
NH_4	Top	7.943	0.98
NH_4	Mid	0.912	0.99
NH_4	Sub	0.380	0.99

so that the sampling variances and covariances of the estimated means can be estimated by:

$$\begin{aligned} \mathrm{Var}_{pj} &= \frac{2\hat{\sigma}_j^2}{n} \\ \hat{\sigma}_j^2 &= \frac{1}{n-1}\sum_{u=1}^{n}\left(y_{ju}-\hat{\bar{y}}_j\right)^2 \\ \mathrm{Cov}_{jk} &= \frac{2\hat{\sigma}_{jk}^2}{n} \\ \hat{\sigma}_{jk}^2 &= \frac{1}{n/2-1}\sum_{u=1}^{n/2}\left(y_{ju}-\hat{\bar{y}}_j^{(m)}\right)\left(y_{ku}-\hat{\bar{y}}_k^{(m)}\right) \end{aligned} \tag{18}$$

with $\hat{\sigma}_j^2$ the estimated spatial variance at time t_j, $\hat{\sigma}_{jk}^2$ the estimated spatial covariance at time t_j and t_k, n the number of sampling locations (MFV plots) per sampling time, $\bar{y}_j$ the sample mean at time t_j, and $\hat{\bar{y}}_j^{(m)}$ the sample mean in the panel sampled at time t_j and t_k. Note that the spatial variance was estimated from two panels with n locations, whereas the spatial covariance could be estimated from one panel with $n/2$ sampling locations only.

We assumed a first order autoregressive model for the spatial means. For this model the matrix $\mathbf{C}_\xi$ equals

$$\begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}\tau^2$$

with ρ the AR parameter (auto-correlation at time lag 1), and τ^2 the model variance of the spatial means (variance of model errors). The combination of ρ and τ^2 with a minimum value for the negative residual loglikelihood was found by computing L_R (Eq. (17)) for a fine grid of 99×402 nodes. Parameter ρ ranged from $0.01, 0.02 \cdots 0.99$ and τ^2 ranged from $0, 10^{-3.00}, 10^{-2.99} \cdots 10^{1.00}$.

Contour plots of the negative residual loglikelihood showed that for none of the variables there was a well defined minimum for the model variance parameters ρ and τ^2 (Fig. 2). For all variables except pH in the middle soil horizon, the contour plots show a flat-bottomed, more or less horizontally oriented gully that sharply bends upward at the right side of the plot. For all variables, except nitrate in the topsoil and in the middle soil horizon, the gully drains to the right, and as a consequence the fitted value of ρ (nearly) equaled the maximum of 0.99. For nitrate in the top soil and middle soil horizon the gully drained to the left and the fitted value of ρ was the minimum of 0.01 (Table 1). However, the bottom of the gullies at the outlet were only marginally deeper than more upstream. Finally, for several variables the fitted values for τ^2 were unrealistically large (Table 1). For these reasons we estimated the parameters ρ and τ^2 by an alternative method. In this method the autoregressive parameter ρ was first estimated as the average of the correlations within the panels. Then the model variance τ^2 was estimated by REML conditional on this estimated autoregressive parameter. For pH at all three depths and ammonium in the subsoil the two elementary estimates of the mean at a given sampling time were first combined to avoid the ill-conditioned covariance matrices.

3.1. Results

The fitted trends were comparable with the trends defined as a population parameter as reported by Brus and de Gruijter (2011) (Table 2, Fig. 3). For nitrate and ammonium in the topsoil the trends differed most. More important is the increase of the standard error compared to the standard error of the estimated trend defined as a population parameter. This is as expected because in the hybrid sampling approach the model error is included whereas this error was ignored in the design-based approach as described by Brus and de Gruijter (2011). The increase in standard error was the largest for nitrate and ammonium in the topsoil, which is in agreement with the large fitted values for the model variance τ^2 for these variables. Due to the increased standard error the trend for ammonium in the subsoil was not significant anymore at the 95% confidence level. The trend of pH in the sub-soil still was significant.

In all cases except pH in the middle soil horizon the contribution of the model variance to the standard error of the estimated trend exceeds the contribution of the sampling variance. This is especially true for the topsoil. For pH the fitted value for τ^2 was zero and consequently the contribution of the model error was zero.

In all cases the sampling standard error of the estimated trend parameter was somewhat smaller than the sampling standard error of the estimated trend defined as a population parameter.

Table 2
Estimated trend ($\hat{\beta}_1$)and its standard error ($\mathrm{se}(\hat{\beta}_1)$) for pH, nitrate and ammonium at three depths in forest soils on the 'Utrechtse Heuvelrug', obtained by REML estimation of the variance of model errors (τ^2) conditional on the autoregressive parameter (ρ) that was estimated as the average of correlations within panels; $\mathrm{se}_P(\hat{\beta}_1)$ and $\mathrm{se}_\xi(\hat{\beta}_1)$ indicate the contributions of the sampling variance of the estimated spatial means and of the model variance of the spatial means to the standard error of the trend; $\hat{b}$ and $\mathrm{se}(\hat{b})$: estimate of trend defined as a population parameter, and its standard error, see Brus and de Gruijter (2011). Numbers for estimated trend in italics: significant at a confidence level of 95%.

Comp.	Depth	ρ	τ^2	$\hat{\beta}_1$	$\mathrm{se}(\hat{\beta}_1)$	$\mathrm{se}_P(\hat{\beta}_1)$	$\mathrm{se}_\xi(\hat{\beta}_1)$	$\hat{b}$	$\mathrm{se}(\hat{b})$
pH	Top	0.75	0.000	0.0284	0.0285	0.0285	0.0000	0.0218	0.0315
pH	Mid	0.63	0.000	0.0195	0.0233	0.0233	0.0000	0.0213	0.0173
pH	Sub	0.66	0.000	*−0.0691*	0.0160	0.0160	0.0000	*−0.0626*	0.0185
NO_3	Top	0.64	0.355	0.0301	0.2485	0.0615	0.2404	−0.0067	0.0652
NO_3	Mid	0.39	0.032	0.0047	0.0936	0.0413	0.0808	0.0037	0.0483
NO_3	Sub	0.52	0.013	0.0201	0.0603	0.0293	0.0502	0.0079	0.0350
NH_4	Top	0.23	0.186	−0.0400	0.2625	0.0712	0.1990	−0.0167	0.1758
NH_4	Mid	0.28	0.010	−0.0789	0.0674	0.0416	0.0461	−0.0864	0.0533
NH_4	Sub	0.44	0.009	0.0321	0.0467	0.0153	0.0434	*0.0368*	0.0165

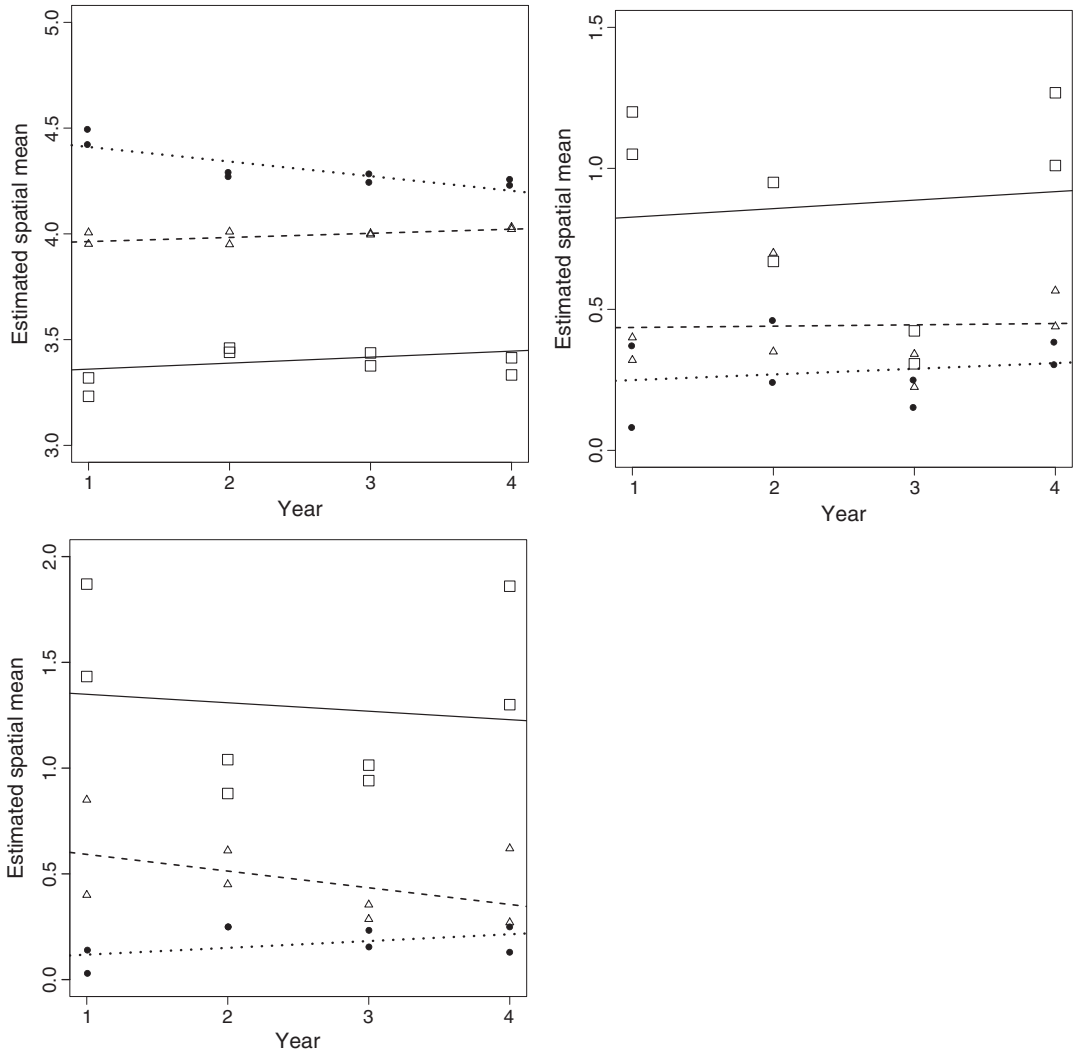

Fig. 3. Estimated trend of spatial means of pH (top-left), nitrate (top-right) and ammonium (bottom-left) in forest soils of Utrechtse Heuvelrug at three depths (squares, solid line: topsoil; triangles, dashed line: middle soil horizon; dots, dotted line: subsoil).

4. Discussion

4.1. Hybrid versus fully model-based approach

In the hybrid approach the errorless spatial means, i.e. the spatial means that would be obtained when the area is sampled exhaustively, are considered as random variables. This implies that also the value at a given location is considered as random. In other words, we are considering a stochastic space–time process. In the fully model-based approach the variation in both space and time is described by a model. In contrast to this fully model-based approach, 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 quantity, two stochastic processes are accounted for, the random selection of the sampling locations *and* the stochastic space–time process.

As for spatial sampling, estimation of global quantities like the mean is generally most efficiently done by a design-based strategy, i.e. by some form of random sampling and inference based on the sampling design (Brus and de Gruijter, 1997). The reason is that for design-based strategies no locations need to be sampled specially for estimating a variogram, because no variogram is needed. For estimating a spatial 2D-variogram by the method-of-moments, Webster and Oliver (1992) recommended the sampling of at least 150 locations. There is some evidence that maximum likelihood estimation of the variogram is more efficient in some situations (Lark, 2000), but still substantial numbers of locations are required. Many of these locations should form pairs at short mutual distances in order to model the short range variation. Such sampling patterns, however, are generally sub-optimal for estimating global means. Apart from efficiency, due to the random selection of locations, design-based strategies offer design-unbiasedness of the estimated mean, a stronger quality criterion than the model-unbiasedness resulting from model-based strategies, which do not pose restrictions on the way the locations are selected.

With regard to sampling in time, purposive selection of sampling times will be more appropriate for the purpose of estimating temporal trends than random selection. The reason is that with purposive selection the sampling times can in principle be chosen such that the accuracy of the estimated trend model parameters is maximized (Myers and Montgomery, 1995). For instance, if three sampling times are to be selected, then it is clearly more efficient to take as sampling times the beginning, the end and half-way the monitoring period than any three randomly selected times.

The fully model-based approach involves a complete space–time model of the variation, whereas in the hybrid approach a much simpler model will do, *viz.* a time-series model of the spatial means.

This simplicity renders the following advantages of the hybrid approach compared with the fully model-based approach.

First, less data are needed. As mentioned above, quite a number of observations is needed already for estimating a 2D-variogram. By adding a third dimension (time), a multiple of this will be needed for estimating a space–time variogram. For modeling the spatio–temporal correlation, clusters of observations in the space–time universe are required, such as high-frequency time-series of observations, spatial transects of observations at short mutual distances and, ideally, clusters of observations crossing both space and time.

Second, because less strong assumptions are needed, mainly on stationarity and anisotropy, the validity of estimation methods based on the hybrid approach will be less easily jeopardized than those based on the fully model-based approach. For the same reason estimation methods based on the hybrid approach are likely more robust than those based on the fully model-based approach, in the sense that the inference results are less sensitive to deviations from the model assumptions.

The hybrid approach presented here can also be of use for estimating other global space–time parameters such as the space–time mean or total, think for instance of the total annual greenhouse gas emissions in a region. Although a fully design-based sampling approach for estimating the space–time mean can be advantageous, in practice researchers are often reluctant to select times by probability sampling. Often sampling times are selected non-randomly at constant interval. Reasons for the reluctance are potential inefficiency and practical problems. In these situations it still can be recommendable to select locations by probability sampling, enabling design-based estimation of spatial means, followed by model-based prediction of the space–time mean or total.

4.2. Definitions of trend and their error variances

In this paper the trend is defined as a model parameter, whereas Brus and de Gruijter (2011) defined the trend as a population parameter:

$$b_2 = \sum_{j=1}^{r} w_j \bar{y}_j \tag{19}$$

with $\bar{y}_j$ the (errorless) spatial mean at time t_j, and w_j the weight attached to the mean at time t_j:

$$w_j = \frac{t_j - \bar{t}}{\sum_{j=1}^{r} \left(t_j - \bar{t}\right)^2} \tag{20}$$

With exhaustive spatial sampling at all selected sampling times, the error variance of the estimated trend defined as in Eq. (19) equals zero; there is no uncertainty left. The reason is that with this definition the universe of interest consists of a finite set of (infinite) spatial populations:

$$U = \{\mathcal{A}_1, \mathcal{A}_2 \cdots \mathcal{A}_r\} \tag{21}$$

with $\mathcal{A}_1$ the spatial population at sampling time t_1, *et cetera*. This universe is a subset only of the entire space–time universe, defined as the Cartesian product of the continuous temporal universe $\mathcal{T}$ and the spatial universe $\mathcal{A}: \mathcal{U} = \mathcal{T} \times \mathcal{A}$. The trend is not a function of the spatial means at times other than the sampling times, so uncertainty about the spatial means at other times does not contribute to the uncertainty about the estimated trend as defined in Eq. (19).

As opposed to this, with exhaustive spatial sampling at the selected sampling times the error variance of the estimated trend defined as a model parameter is larger than zero. This is even true with exhaustive sampling of the entire space–time universe. The reason is that both sampling error and model error contribute to the error variance, and even with exhaustive sampling of the space–time universe there would still be a contribution of the model error.

A third option is to define the trend as (ter Braak et al., 2008):

$$b_2 = \frac{\int_{T_s}^{T_e} (t - \bar{t}) \bar{Y}_t \mathrm{d}t}{\int_{T_s}^{T_e} (t - \bar{t})^2 \mathrm{d}t} \tag{22}$$

with T_s and T_e the start and the end of the temporal universe $\mathcal{T}$, and $\bar{t} = \frac{T_s + T_e}{2}$. With this definition the trend is a population parameter again, but the universe of interest now is the entire space–time universe $\mathcal{A}: \mathcal{U} = \mathcal{T} \times \mathcal{A}$. ter Braak et al. (2008) give the BLUP for this trend parameter b_2 under a fully model-based sampling approach. Contrary to the trend defined as in Eq. (19) with exhaustive spatial sampling at the selected sampling times the standard error of the predicted trend will be larger than zero. The reason is that we are uncertain about the spatial means at the other times. Unlike the trend defined as a model parameter, with exhaustive sampling of the space–time universe the error variance of the predicted trend defined as a population parameter equals zero, because only sampling error is now involved.

Finally, the trend can be defined as a parameter of a time-series model for the entire temporal universe $\mathcal{T}$, not just for a finite set of times as in Eq. (2). In this case the model parameter can again be estimated by the fully model-based approach of ter Braak et al. (2008).

Summarizing the temporal trend of spatial means can be defined as a population parameter or as a model parameter, and both definitions may be applied to a discrete or a continuous temporal universe of interest. This leads to four combinations:

1. trend as population parameter with a discrete temporal universe of interest. Brus and de Gruijter (2011) discuss design-based inference for this case.
2. trend as population parameter with a continuous temporal universe of interest. ter Braak et al. (2008) discuss model-based inference for this case.
3. trend as model parameter with a discrete temporal universe of interest. In this paper we propose a hybrid, design-based and model-based approach.
4. trend as model parameter with a continuous temporal universe of interest. See ter Braak et al. (2008) for model-based inference for this case.

For any given monitoring project one has to decide which one of the above combinations should be applied. The choice between a discrete or a continuous temporal universe should be guided by the ultimate purpose of the monitoring. If this purpose implies that the spatial means or totals at all times in the monitoring period are of interest, then clearly one has to deal with a continuous temporal universe of interest. Otherwise it will be a discrete temporal universe. Examples of the latter are when interest is in the nutrient status of the soil at the start of the growing season or in the maxima and minima of soil biota with cyclic variation of population densities. Also, with only a few sampling times, estimation of the trend defined on a continuous temporal universe of interest can be unfeasible, so that we are forced to adopt a a discrete temporal universe.

The choice between trend as a population parameter or as a model parameter should also be guided by the purpose of the monitoring, but here the matter seems to be more complicated. As a provisional indication, we would generally opt for a definition as population parameter if the trend will play a role in some form of status monitoring or compliance monitoring. On the other hand, if the trend has to be used in forecasting, then we would advise a definition as model parameter, because then one would expect more realistic prediction intervals.

4.3. Spatial sampling design

The sampling approach proposed here aims at estimating spatial means (totals, proportions) at the selected times, and the evolution over time of these spatial means. In practice sponsors of sampling such as governments and regulators, are often a whole lot fuzzier about what they want than we might like them to be. It often happens that they both want to estimate spatial means for the area as a whole or for several subareas (soil-landscape units) and mapping. In this situation we recommend a probability sampling design that leads to optimal coverage of geographical space. Examples are systematic random sampling or random sampling from compact geographical strata formed by k-means (Brus et al., 1999; Walvoort et al., 2010). For the former design no unbiased estimators of the sampling variance exists, whereas for the latter design such estimator does exist when at least two points per geostratum are selected.

5. Conclusions

In monitoring the soil the building of a model that describes the variation of the soil in both space and time generally is quite challenging. When the aim is to monitor spatial means or totals, the building of such a space–time model can be avoided by random selection of the sampling locations. Probability sampling in space enables model-free, design-based estimation of spatial means. The temporal trend of the spatial means and other global space–time parameters such as the space–time mean, can then be estimated by building a time-series model of the spatial means. Such a model is much simpler than a complete space–time model as used in the fully model-based approach. This leads to a hybrid, design-based and model-based sampling approach which is described in this paper. In this hybrid approach both the sampling error and the model error are accounted for in quantifying the uncertainty about the monitoring result.

We think that the hybrid design-based model-based sampling approach can be an attractive alternative to a fully model-based approach for estimating the temporal trend of spatial means and other global space–time parameters. By only partly modeling the space–time process less data are required. Besides, the validity of estimation methods based on the hybrid approach will be less easily jeopardized than those based on the fully model-based approach because less stronger assumptions are needed in the hybrid approach. For the same reason estimation methods based on the hybrid approach are likely more robust than those based on the fully model-based approach, in the sense that the inference results are less sensitive to deviations from the model assumptions.

In a case study on soil acidification and eutrophication the estimated linear trends under the hybrid approach were comparable to the estimated trends under the design-based approach (Brus and de Gruijter, 2011). However, for most variables their standard errors were larger, especially for nitrate and ammonium in the topsoil, which showed strong model fluctuations around the linear trend. This is as expected because in the hybrid sampling approach the model error is included whereas this error was ignored in the design-based approach.

The temporal trend of spatial means can be defined either as a population parameter or as a model parameter, and both definitions may be applied either to a discrete or a continuous temporal universe of interest. This leads to four definitions. The choice of definition should be guided by the purpose of monitoring. If the spatial means (totals) at all times in the monitoring period are of interest then a trend should be defined on a continuous temporal universe of interest. If interest is in the status of the soil at specific times, for instance the start of the growing season, then the trend should be defined on a discrete time domain. Also, with only a few sampling times, estimation of the trend defined on a continuous temporal universe of interest can be unfeasible, so that we are forced to adopt a a discrete temporal universe. For status and compliance monitoring defining the trend as a population parameter is more appropriate, whereas for forecasting a definition as a model parameter is more appropriate as it leads to more realistic prediction intervals.

References

Binder, D.A., Hidiroglou, M.A., 1988. Sampling in time. In: Krishnaiah, P.R., Rao, C.R. (Eds.), Handbook of Statistics, volume 6, pp. 187–211. North-Holland, Amsterdam.

ter Braak, C.J.F., Brus, D.J., Pebesma, E.J., 2008. Comparing sampling patterns for kriging the spatial mean temporal trend. Journal of Agricultural, Biological, and Environmental Statistics 13, 159–176.

Brus, D.J., de Gruijter, J.J., 1993. Design-based versus model-based estimates of spatial means. Theory and application in environmental soil science. Environmetrics 4, 123–152.

Brus, D.J., de Gruijter, J.J., 1997. Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with discussion). Geoderma 80, 1–59.

Brus, D.J., de Gruijter, J.J., 2011. Design-based Generalized Least Squares estimation of status and trend of soil properties from monitoring data. Geoderma 164, 172–180.

Brus, D.J., Heuvelink, G.B.M., 2007. Optimization of sample patterns for universal kriging of environmental variables. Geoderma 138, 86–95.

Brus, D.J., Knotters, M., 2008. Sampling design for compliance monitoring of surface water quality: a case study in a polder area. Water Resources Research 44, W11410.

Brus, D.J., Spätjens, L.E.E.M., de Gruijter, J.J., 1999. A sampling scheme for estimating the mean extractable phosphorus concentration of fields for environmental regulation. Geoderma 89, 129–148.

Diggle, P., Ribeiro Jr., P., 2007. Model-based Geostatistics. Springer.

van Groenigen, J.W., Siderius, W., Stein, A., 1999. Constrained optimisation of soil sampling for minimisation of the kriging variance. Geoderma 87, 239–259.

de Gruijter, J.J., ter Braak, C.J.F., 1990. Model-free estimation from spatial samples: a reappraisal of classical sampling theory. Mathematical Geology 22, 407–415.

de Gruijter, J.J., Brus, D.J., Bierkens, M.F.P., Knotters, M., 2006. Sampling for Natural Resource Monitoring. Springer, Berlin.

Jones, R.G., 1980. Best linear unbiased estimators for repeated surveys. Journal of the Royal Statistical Society. Series B (Methodological) 42, 221–226.

Lark, R.M., 2000. Estimating variograms of soil properties by the method-of-moments and maximum likelihood. European Journal of Soil Science 51, 717–728.

Lark, R.M., Cullis, B.R., Welham, S.J., 2006. On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor E-BLUP with REML. European Journal of Soil Science 57, 787–799.

Myers, R.H., Montgomery, D.C., 1995. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley, New York.

Papritz, A., Webster, R., 1995. Estimating temporal change in soil monitoring: I. Statistical theory. European Journal of Soil Science 46, 1–12.

Walvoort, D.J.J., Brus, D.J., de Gruijter, J.J., 2010. An R package for spatial coverage sampling and random sampling from compact geographical strata by k-means. Computers and Geosciences 36, 1261–1267.

Webster, R., Oliver, M.A., 1992. Sample adequately to estimate variograms of soil properties. Journal of Soil Science 43, 177–192.