Uncertainty of variance component estimates in nested sampling: a case study on the field-scale spatial variability of a restored soil

Uncertainty of variance component estimates in nested sampling: a case study on the field-scale spatial variability of a restored soil [1]

A. Papritz[a], A. Dümig[b], C. Zimmermann[c], H. H. Gerke[d], B. Felderer[a], I. Kögel-Knabner[b], W. Schaaf[c] & R. Schulin[a]

[a]*Institute of Terrestrial Ecosystems, ETH Zurich, CH-8092 Zürich, Switzerland,* [b]*Lehrstuhl für Bodenkunde, TU München, D-85350 Freising-Weihenstephan,* [c]*Soil Protection and Recultivation, Brandenburg University of Technology Cottbus, D-03013 Cottbus and* [d]*Institute of Soil Landscape Research, Leibniz Centre for Agricultural Landscape Research (ZALF) Müncheberg, D-15374 Müncheberg, Germany*

Running head: *Uncertainty of variance components in nested sampling*

Correspondence: A. Papritz. E-mail: papritz@env.ethz.ch

Summary

We studied the variation of soil properties on a six-ha artificial catchment constructed near Cottbus, Germany, to investigate processes of initial ecosystem genesis. We wanted to see whether spatial auto-correlation patterns could be identified three years after site construction. Topsoil was sampled at 192 locations using a balanced nested design involving six spatial scales (0.2 m to >60 m) and analysed for particle size, organic matter content, pH, soluble P, and various fractions of selected metals. Variance components were estimated by residual maximum likelihood. The uncertainty of variance estimates was characterized by the Fisher Information matrix and likelihood joint confidence regions. The latter approach was used for the first time to characterize uncertainties of variance estimates in spatial nested sampling. Likelihood ratio tests showed that all variables were spatially auto-correlated, but the allocation of the variance to specific spatial scales was highly uncertain. For most variables, at least one variance component could not be estimated precisely because the profile likelihood was either flat or the maximum lay on the boundary of the parameter space. Uncertainty estimates derived from the Fisher Information either could not be computed or were unrealistic in these cases. Likelihood joint confidence regions gave more realistic uncertainty estimates. Joint confidence regions for accumulated variance components showed that the shape of the estimated variograms was poorly defined for most variables. Simulations indicated that poor identification of variance components might be a general problem of nested sampling surveys, which has been under-estimated in the past. Hence, our work provides some incentive for re-examining the statistical properties of the methodology.

Introduction

Biotic and abiotic processes cause characteristic spatial (and temporal) patterns in terrestrial ecosystems. Studying these processes is one way to learn about those dominating energy and matter fluxes and interactions between organisms within an ecosystem. Soil scientists and ecologists have long since recognized this, but "an explicit focus on understanding spatial heterogeneity – revealing its myriad abiotic and biotic causes and its ecological consequences – emerged in the 1980s as landscape ecology developed and spatial data and analysis methods became more widely available" (Turner, 2005).

One method to investigate spatial patterns is the analysis of variance of data collected in surveys in which the sampled locations are spatially nested. In soil science this nested sampling approach was popularized by Oliver & Webster (1986, 1987; see also Webster & Oliver, 2007). Originally, the variance components associated with each level of a nested design were estimated from the mean

[1]Published in *European Journal of Soil Science*, Volume **62**: 479–465. The definite version is available at `http://www.blackwell-synergy.com`

squares of analysis of variance tables. More recently, likelihood based estimation methods (Pinheiro & Bates, 2000) gained acceptance (Webster *et al.*, 2006). Through these, the scope of nested sampling could be considerably broadened (Lark, 2005; Corstanje *et al.*, 2007; Lark & Corstanje, 2009), as the approach thus became incorporated in linear mixed modelling methodology (Pinheiro & Bates, 2000; Jiang, 2007).

We used nested sampling to elucidate the spatial scales of variation of chemical and physical soil properties in a six-ha catchment that was artificially constructed in the Lusatian lignite mining landscape near Cottbus, NE Germany. The over-arching hypothesis of the consortium of research groups studying the development of this catchment was that "initial patterns define and shape the development and later stages of an ecosystem" (Gerwin *et al.*, 2009). Spatial patterns are assumed to evolve in the course of ecosystem development as the result of processes such as erosion, formation of preferential drainage paths on the surface and within the soil, non-uniform chemical weathering or uneven colonization by organisms.

In order to detect the formation of spatial patterns, a comprehensive characterization of the initial state of the soil and of its variation across the catchment was a prerequisite. For this purpose, two surveys were conducted: first after completion of construction a survey in which soil was sampled at two depths (0 – 30 cm, 30 – 100 cm) on a 20-m × 20-m grid (Gerwin *et al.*, 2009), and in summer 2008 a nested sampling survey, in which only the topsoil (0 – 3 cm) was sampled. The second survey was delayed for reasons not under our control. This is unfortunate, because water and wind erosion have shaped the soil surface since autumn 2005 (Gerwin *et al.*, 2009) and plants quickly started to colonize the catchment. Thus, the nested sampling survey did not capture the state of the ecosystem at time zero, although still at a very early stage of its development.

Here, we report a detailed analysis of the data collected by the nested sampling campaign in 2008. The survey had the following objectives: We wanted to explore whether selected physical and chemical topsoil properties *varied in a purely random way* across the catchment or whether they showed *spatially structured* variation (auto-correlation). In the latter case, we were interested to *estimate the scales of variation*, i.e. the distances over which the measurements were spatially dependent, and to compare the estimated scales with recorded patterns of the construction process. We used maximum likelihood methods for estimating the variance components associated with the levels of the nested design. In the course of the analyses, we discovered that rather often we could determine the magnitude of the variance components with only unsatisfactory precision. Therefore, we complemented the statistical analyses by performing simulations which suggested that the poor precision of variance estimates might be a general problem of the nested sampling methodology. Cole (2009) remarked that uncertainties of variance component estimates have rarely been reported in the ecological literature. Here, we investigate the uncertainty of variance component estimates and the related identification problems for the first time in an application of nested sampling in soil science. We are currently aware of only one study (Rawlins *et al.*, 2009) which has reported standard errors of variance component estimates.

Material and methods

'Chicken Creek' catchment

The artificial catchment was constructed in 2004 – 2005 by the company Vattenfall Europe Mining AG in collaboration with the Brandenburg University of Technology Cottbus as headwaters of a stream named "Chicken Creek". It is located SE of the city of Cottbus, NE Germany, in an area where lignite is extracted by open-cast mining. Kendzia *et al.* (2008) and Gerwin *et al.* (2009) describe the site and the construction of the catchment. Here we summarize the aspects which are important for our analysis.

The catchment was constructed from coarse-textured over-burden sediments (sand, sandy loam) of the Pleistocene (terminal moraine of Wolstonian stage). This material was removed by bucket wheel excavators from the outcrop side of the open-cast mine and delivered by conveyer belts to stackers that deposited it on a clay layer that seals the catchment at its base. During this process the material was deposited in form of slightly bent and elongated spoil heaps, looking from above like rows of connected cones (Figure 1, inset A). The majority of these heaps were 20 – 50 m long and 2 – 5 m wide. At the points where the cantilever started to swing back the heaps became wider (up to 10 m wide, Figure 1, inset B). The heaps were then levelled by pushing material from the ridges

into the voids between them (Figure 1, inset C). Thereafter the surface was further graded by pulling rails over the terrain. No further amelioration (fertilizer addition, liming, erosion control) took place. The material of the SW and NE parts of the catchment was not deposited at the same time (NE: July 2004, SW: September 2004). Since its properties varied with the location from where it was taken and since it was not stored and mixed, the SW and NE part of the catchment were not built from exactly the same material. Some heterogeneity in the dumped substrate is visible in Figure 1 as differences in the grey shading. Since the completion of the construction works in September 2005, human interferences have been kept to a minimum. The catchment has an area of about 6 ha and faces SE with a mean slope of about 3.5% along its main axis (Figure 2, left panel).

Nested sampling survey

A total of $N = 192$ locations for soil sampling were selected using a balanced nested design (Webster & Oliver, 2007). Available resources and the wish to minimize disturbances by soil sampling limited the sample size.
At the first level of the nested classification, three pairs of clusters were purposively selected in the SW and NE part of the catchment, each cluster consisting of 16 locations. Two pairs of clusters lay in the upper part of the back slope area, two in its middle and two in the transition zone to the steeper foot slope area (Figure 2, left panel). The distance between two clusters of a pair was 20 m. This corresponds to the mesh width of the 2005/2006 grid survey (Gerwin *et al.*, 2009).
The grouping of locations within each cluster is schematically depicted in Figure 2 (right panel): There were two groups of eight locations (octuples), separated by a distance of 6 m, each consisting in turn of two quadruples, spaced 2 m apart. The locations within quadruples were grouped into doubles, 0.6 m away from each other. Finally, the two locations of a double were separated by 0.2 m. This resulted in a total of 6 cluster pairs, 12 clusters, 24 octuples, 48 quadruples and 96 doubles. The spatial arrangement of octuples, quadruples, doubles was the same within all clusters.
The observations of a response variable, Y_{ijklmn}, collected by this design, can be represented by a linear mixed-effects model (Pinheiro & Bates, 2000):

$$Y_{ijklmn} = \mathbf{x}_{ijklmn}^{\mathrm{T}}\,\boldsymbol{\beta} + P_i + C_{ij} + O_{ijk} + Q_{ijkl} + D_{ijklm} + \varepsilon_{ijklmn}, \tag{1}$$

where $^{\mathrm{T}}$ denotes transpose; $\mathbf{x}_{ij\ldots n}$ and $\boldsymbol{\beta}$ are p-vectors of covariates and regression coefficients representing the fixed effects; $\varepsilon_{ij\ldots n}$ is the residual error ($n = 1, 2$); and the remaining terms denote the nested random effects for cluster pairs ($P_i, i = 1, 2, \ldots, 6$; spatial scale >60 m), clusters ($C_{ij}, j = 1, 2$; 20 m), octuples ($O_{ijk}, k = 1, 2$; 6 m), quadruples, ($Q_{ij\ldots l}, l = 1, 2$; 2 m), and doubles, ($D_{ij\ldots m}, m = 1, 2$; 0.6 m). The random effects and the residual errors were assumed to be normally distributed with zero mean and variances $\sigma^2_{>60\mathrm{m}}$ (P_i), $\sigma^2_{20\mathrm{m}}$ (C_{ij}), $\sigma^2_{6\mathrm{m}}$ (O_{ijk}), $\sigma^2_{2\mathrm{m}}$ ($Q_{ij\ldots l}$), $\sigma^2_{0.6\mathrm{m}}$ ($D_{ij\ldots m}$) and $\sigma^2_{0.2\mathrm{m}}$ ($\varepsilon_{ij\ldots n}$). At a given level of the design, the random effects were further assumed to be independent for different indices, independent of the random effects of other levels and of the residual errors. The number of degrees of freedom (df) associated with the various levels of this balanced nested design are listed in Table 1 (first row).
Three points are worth noting: firstly, Equation (1) implies (see also Figure 2, right panel) that an observation is modelled as the sum of the spatial trend, plus the random deviation of the mean of a cluster pair from this trend, plus the random difference of the cluster and the cluster pair means, etc. Secondly, the variance of the observations is:

$$\mathrm{Var}\,[Y_{ij\ldots n}] = \sigma^2_{0.2\mathrm{m}} + \sigma^2_{0.6\mathrm{m}} + \ldots + \sigma^2_{>60\mathrm{m}}, \tag{2}$$

by virtue of the independence of the random effects across different levels. Thirdly, the auto-correlation between two observations is equal to the intraclass correlation (ratio of the sum of variances of the 'shared' random effects to $\mathrm{Var}\,[Y_{ij\ldots n}]$, compare for example Pinheiro & Bates, 2000). As an example, the correlation of two observations within a given octuple, say Y_{ijklmn} and $Y_{ijkl'm'n'}$, $l \neq l'$, is equal to:

$$\mathrm{Corr}\,[Y_{ijklmn}, Y_{ijkl'm'n'}] = \frac{\sigma^2_{6\mathrm{m}} + \sigma^2_{20\mathrm{m}} + \sigma^2_{>60\mathrm{m}}}{\sigma^2_{0.2\mathrm{m}} + \sigma^2_{0.6\mathrm{m}} + \ldots + \sigma^2_{>60\mathrm{m}}}. \tag{3}$$

Clearly, the correlation increases with an increasing number of 'shared' random effects. In spatial sampling this means that (i) the auto-correlation increases with decreasing distance between two

sampled locations and that (ii) there is no spatial dependence if all variances except $\sigma^2_{0.2\mathrm{m}}$ vanish. The semivariance, $\gamma(\cdot)$, is usually preferred over the correlation or covariance to characterize auto-correlation. In nested sampling, the semivariance of two observations is equal to the sum of the variances of the 'non-shared' random effects (Miesch, 1975; Webster & Oliver, 2007). Since the average spacing between two locations within an octuple is equal to about two metres (Figure 2, right panel), we obtain for our example:

$$\gamma(2\ \mathrm{m}) \approx \frac{1}{2}\mathrm{Var}\left[Y_{ijklmn} - Y_{ijkl'm'n'}\right] = \sigma^2_{0.2\mathrm{m}} + \sigma^2_{0.6\mathrm{m}} + \sigma^2_{2\mathrm{m}}. \quad (4)$$

Field work and laboratory analyses

The sampling points were located in the field by tape measure from the nodes of a 20-m × 20-m grid of reference points that was set up at the outset of the project. Sampling took place in August 2008. The two cores (height 3 cm, volume 40 cm^3) collected at each location next to each other were bulked and filled into plastic bags. In the laboratory the soil samples were dried for 48 hours at 40° C, plant fragments were manually removed, and the soil was passed through 2-mm sieves.
For the analysis of the particle size distribution, subsamples were pretreated with H_2O_2 at 80° C overnight to destroy the organic matter. We did not dissolve carbonates by an acid pre-treatment. The remaining mineral soil was dispersed in 6 ml of a 0.2% sodium-hexametaphosphate solution by overhead shaking for two hours, followed by one minute exposure to low energy (15 kHz, 75 W) ultrasonic vibration. The particle size distribution of the resulting suspension was analysed for the size range 0.04 μm – 2 mm by using laser diffractometry (LS13320 instrument, Beckman Coulter Inc., Brea CA, USA). The following particle size fractions were used in the statistical analyses: clay (0.04 – 2 μm), silt (2 – 63 μm), sand (63 – 630 μm) and coarse sand (0.63 – 2 mm).
Soil pH was measured in a 1:2.5 water-to-soil suspension using a pH-electrode (SenTix41 electrode, WTW Wissenschaftlich-Technische Werkstätten GmbH, Weilheim, Germany). The organic matter content (OM) was determined as the weight loss after combustion at 430° C in a muffle furnace. Oxalate extractable metal concentrations ($\mathrm{Fe_o}$, $\mathrm{Mn_o}$, $\mathrm{Al_o}$) were determined using the procedure of Schwertmann (1964). Total Fe in oxides ($\mathrm{Fe_d}$) was determined by a dithionite-citrate extraction using the method of Mehra & Jackson (1960). Soluble phosphorus ($\mathrm{P_{re}}$) was estimated from resin extractable P (Saggar *et al.*, 1990). Phosphorus concentrations in solution were determined photometrically according to Van Veldhoven & Mannaerts (1987). The total concentrations of K, Ca, Fe, Mn and Al in the soil samples were measured by XRF spectroscopy using a Spectro X-lab 2000 instrument (SPECTRO Analytical Instruments GmbH, Kleve, Germany). These measurements are missing for four soil samples because a first analysis failed and we had not enough soil sample to repeat this.

Statistical analyses

We used the software environment *R* (R Development Core Team, 2009) with the add-on packages 'robustbase' (computation of robust summaries) and 'nlme' (linear mixed-effects modelling, Pinheiro & Bates, 2000). For data exploration, we employed robust algorithms: robust standard deviations were computed by the Qn-estimator (Rousseeuw & Croux, 1993). Using these as fixed dispersion parameters, we estimated the means robustly using a Huber M-estimate (Maronna *et al.*, 2006, sec. 2.6.1) with the Huber constant $k = 1.5$.
The variance components were estimated by residual maximum likelihood (REML, also called restricted maximum likelihood; Pinheiro & Bates, 2000; Webster *et al.*, 2006). Models were fitted by the *R* function *lme*, which maximizes the residual log-likelihood unconstrainedly with respect to the logarithms of the variance components (Pinheiro & Bates, 2000), and thus always returns positive variance estimates, say $\widehat{\sigma}^2_{0.2\mathrm{m}}, \widehat{\sigma}^2_{0.6\mathrm{m}}, \ldots, \widehat{\sigma}^2_{>60\mathrm{m}}$. For comparison, we estimated the variance components also by our own REML code that maximizes the residual log-likelihood unconstrainedly with respect to the untransformed variances. The latter estimates, say $\widetilde{\sigma}^2_{0.2\mathrm{m}}, \widetilde{\sigma}^2_{0.6\mathrm{m}}, \ldots, \widetilde{\sigma}^2_{>60\mathrm{m}}$, may become negative, which is permissible as long as the covariance matrix of the data remains positive definite (Webster *et al.*, 2006). Webster *et al.* argue that negative variance components arise in spatial nested sampling "if there is some underlying regular feature in the landscape, such as ancient ploughing patterns".

Initially, we used terrain attributes (gradient, curvature, etc.) derived from a laser scan of the ground surface of the catchment recorded in April 2009 (Figure 2, left panel) as fixed effect covariates. However, none of the response variables showed any significant dependence on these covariates. A factor for the SW and NE parts of the catchment was the only fixed effect covariate that we eventually used. We fitted the model (1) first to the untransformed response variables and then to transforms of these. The transformations were selected on the basis of customary residual diagnostic plots (Pinheiro & Bates, 2000). In more detail, we plotted the Pearson residuals against predictions of doubles ($\widehat{D}_{ij\ldots m}$), quadruples ($\widehat{Q}_{ij\ldots l}$), ..., cluster pairs ($\widehat{P}_i$) to check the assumption of homoscedastic $\varepsilon_{ij\ldots n}$, and we used normal quantile-quantile plots to verify that random effects and residual errors were approximately normally distributed. Transformation to natural logarithms was found to be appropriate except for soil texture and OM that were transformed by arcsine $(\sqrt{\cdot})$. As well as the 'full' model (cf. Equation 1) we also fitted 'reduced' models by omitting some random effects. In these cases, the number of groups was expanded accordingly at the next level of the nested design. For example, if C_{ij} was omitted, we used $k = 1, 2, 3, 4$ instead of $k = 1, 2$ at the level of the octuples O_{ik}. We fitted all 32 models with up to six levels of nested random effects and retained the model that minimized the Akaike information criterion (AIC, Pinheiro & Bates, 2000, sec. 2.4). Likelihood ratio tests (e.g. Pinheiro & Bates, 2000; Webster *et al.*, 2006; Lark & Corstanje, 2009) were used to see whether the 'best-fit reduced' model fitted the data less well than the 'full' model. We did not use adjusted versions of the likelihood ratio tests as suggested by Stram & Lee (1994) for testing the significance of variance components. Thus, our test results are conservative in the sense that we have occasionally favoured the alternative ($\sigma^2 > 0$) over the null hypothesis ($\sigma^2 = 0$).
Confidence intervals for the estimated variance components were computed in two ways, firstly, we used the asymptotic multivariate normal distribution of the REML estimates as a basis. To compute confidence intervals from the output of *lme*, Pinheiro & Bates (2000) assume that $\ln \widehat{\sigma}^2_{0.2\mathrm{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\mathrm{m}}$ follow an asymptotic Gaussian distribution. For our own REML code, we adopted the same assumption for $\widetilde{\sigma}^2_{0.2\mathrm{m}}, \ldots, \widetilde{\sigma}^2_{>60\mathrm{m}}$. In either case, the covariance matrix of the (logarithms of the) estimated variances was computed by inverting the observed Fisher Information matrix $\mathbf{J}$. The function *lme*, as an example, computes the covariance matrix of $\ln \widehat{\sigma}^2_{0.2\mathrm{m}}$, ..., $\ln \widehat{\sigma}^2_{>60\mathrm{m}}$ by inverting $\mathbf{J}(\ln \widehat{\sigma}^2_{0.2\mathrm{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\mathrm{m}})$, i.e., the negative Hessian of the residual log-likelihood function with respect to $\ln \sigma^2_{0.2\mathrm{m}}, \ldots, \ln \sigma^2_{>60\mathrm{m}}$ and evaluated at $\widehat{\sigma}^2_{0.2\mathrm{m}}, \ldots, \widehat{\sigma}^2_{>60\mathrm{m}}$. $\mathbf{J}$ thus characterizes the curvature of the log-likelihood surface at its maximum $L_{\max} = L(\widehat{\sigma}^2_{0.2\mathrm{m}}, \ldots, \widehat{\sigma}^2_{>60\mathrm{m}})$, and it measures how strongly 'peaked' the surface is at the REML estimate.
Secondly, we computed joint confidence regions for several variance components simultaneously on the basis of the likelihood ratio test. To do so, we had to compute the residual profile log-likelihood for model (1). The residual profile log-likelihood, say $L_{\mathrm{p}}(\sigma^2_{i_1}, \sigma^2_{i_2}, \ldots, \sigma^2_{i_q})$, of a set of q variance components, $\sigma^2_{i_1}$, $\sigma^2_{i_2}$, ..., $\sigma^2_{i_q}$, is obtained by maximizing the residual log-likelihood for given $\sigma^2_{i_1}$, $\sigma^2_{i_2}$, ..., $\sigma^2_{i_q}$ with respect to the remaining variances, σ^2_j, $j \notin (i_1, \ldots, i_q)$. The likelihood ratio test states that under the null hypothesis $\sigma^2_{i_1} = c_1$, ..., $\sigma^2_{i_q} = c_q$, the difference:

$$2[L_{\max} - L_{\mathrm{p}}(\sigma^2_{i_1} = c_1, \ldots, \sigma^2_{i_q} = c_q)], \tag{5}$$

follows a χ^2_q-distribution with q df. Thus, q-tuples, $\sigma^2_{i_1}$,...,$\sigma^2_{i_q}$, satisfying the inequality:

$$2[L_{\max} - L_{\mathrm{p}}(\sigma^2_{i_1}, \ldots, \sigma^2_{i_q})] \leq \chi^2_q(1-\alpha), \tag{6}$$

where $\chi^2_q(1-\alpha)$ is the $(1-\alpha)$-quantile of the χ^2_q-distribution, form a joint $(1-\alpha)$-confidence region for $\sigma^2_{i_1}, \ldots, \sigma^2_{i_q})$. Of course, in the case of equality, the q-tuples lie on the boundary of this region. By accumulating the variances of q-tuples satisfying 6, starting with the variance component of the smallest spatial scale and adding step-by-step the components of the next smallest scales, a joint confidence region for the semivariances associated with the levels of the nested design, and thereby a confidence region for an approximate variogram, is obtained.
We used Equation (6) to compute univariate confidence intervals of the six variance components. In more detail, we used bisecting to compute the roots of:

$$2[L_{\max} - L_{\mathrm{p}}(\sigma^2_j)] = \chi^2_1(1-\alpha), \tag{7}$$

where $j \in (0.2\text{m}, \ldots, >60\text{m})$ and σ_j^2 was constrained to be positive. For clay, Al, Ca and pH we also computed 6-tuples on the boundary of the 6-dimensional joint confidence region of $\sigma^2_{0.2\text{m}}, \ldots, \sigma^2_{>60\text{m}}$. To this end, we placed a grid into a hyper-rectangle in the 5-dimensional parameter space spanned by $\sigma^2_{0.6\text{m}}, \ldots, \sigma^2_{>60\text{m}}$. The grid had 20 nodes in each direction, resulting in $20^5 = 3.2 \cdot 10^6$ nodes in total. For each grid node (with fixed values for $\sigma^2_{0.6\text{m}}$, ..., $\sigma^2_{>60\text{m}}$) we then computed the roots of:

$$2[L_{\max} - L_{\text{p}}(\sigma^2_{0.2\text{m}}, \sigma^2_{0.6\text{m}}, \ldots, \sigma^2_{>60\text{m}})] = \chi^2_6(1-\alpha), \tag{8}$$

with respect to $\sigma^2_{0.2\text{m}}$. The resulting 6-tuples lie on the boundary of the joint confidence region of $\sigma^2_{0.2\text{m}}, \ldots, \sigma^2_{>60\text{m}}$. Further details are given in the appendix.
We simulated data from the model in Equation (1) to compare the precision of variance component estimates obtained from balanced and various unbalanced nested designs. To simulate one realization of the 32 values of one cluster pair (see Figure 2, right panel), we drew one random number, p_i, from the normal distribution $\mathcal{N}(0, \sigma^2_{>60\text{m}})$. Then we drew two random numbers, c_{i1}, c_{i2}, independently from $\mathcal{N}(0, \sigma^2_{20\text{m}})$, four numbers, $o_{i11}, o_{i12}, o_{i21}, o_{i22}$, from $\mathcal{N}(0, \sigma^2_{6\text{m}})$, eight numbers, $q_{i111}, q_{i112}, q_{i121}, \ldots, q_{i222}$ from $\mathcal{N}(0, \sigma^2_{2\text{m}})$, 16 numbers, $d_{i1111}, \ldots, d_{i2222}$ from $\mathcal{N}(0, \sigma^2_{0.6\text{m}})$, and finally 32 residual errors, $e_{i11111}, \ldots, e_{i22222}$, from $\mathcal{N}(0, \sigma^2_{0.2\text{m}})$. The 32 simulated observations, $y_{i11111}, \ldots, y_{i22222}$ were then obtained from:

$$y_{ijklmn} = p_i + c_{ij} + o_{ijk} + q_{ijkl} + d_{ijklm} + e_{ijklmn}. \tag{9}$$

A simulated data set comprised 32 independently simulated cluster pairs, each consisting of 32 simulated values, yielding 1024 observations y_{ijklmn}, $i = 1, 2, \ldots, 32$; $j, k, l, m, n \in (1, 2)$ in a data set. In practice, we simulated 1000 such data sets from the variance components estimated for Ca ($\widehat{\sigma}^2_{0.2\text{m}} = 0.0654$, $\widehat{\sigma}^2_{0.6\text{m}} = 0.0082$, $\widehat{\sigma}^2_{2\text{m}} = 0.0379$, $\widehat{\sigma}^2_{6\text{m}} = 0.0158$, $\widehat{\sigma}^2_{20\text{m}} = 0.0179$, $\widehat{\sigma}^2_{>60\text{m}} = 0.0819$). For some designs, less than 32 cluster pairs were needed. In these instances, we selected the required number of cluster pairs randomly from a data set. Furthermore, the unbalanced designs did not include all 32 values of a cluster pair. Again, we selected the required observations randomly within a cluster pair. We fitted the model with a fixed effect for the SW/NE contrast, although we did not add any such effect in the simulations. The statistical properties of the estimated variance were characterized by the bias:

$$\text{BIAS} = \frac{1}{1000} \sum_{i=1}^{1000} \widehat{\sigma}_i^2 - \sigma^2, \tag{10}$$

and the root mean square error:

$$\text{RMSE} = \sqrt{\frac{1}{1000} \sum_{i=1}^{1000} (\widehat{\sigma}_i^2 - \sigma^2)^2}, \tag{11}$$

where σ^2 denotes an arbitrary variance component and $\widehat{\sigma}_i^2$ its estimate computed from the ith simulated data set.

Results and discussion

Exploratory data analysis

Figure 3 shows the histograms of the observations for selected variables, along with kernel probability density estimates, computed separately for the SW and NE parts of the catchment. The histograms of some variables were either bimodal (OM, Fe_o, Al) or platykurtic (sand, K, Fe, P_re). The density curves differed for most variables between the SW and NE parts of the catchment (possible exceptions are pH and Mn). The substrate of the NE part had more sand and P_re, and less clay, silt (not shown), OM, Fe and Al. The SW/NE differences in the sand and OM content agree with results reported by Gerwin *et al.* (2009).
Figure 4 displays the spatial distribution of the measurements for selected variables. The geographical coordinates of the sampling points were rotated and shifted to allow for observing individual values. The differences between the SW (clusters $x2, x3$) and NE parts (clusters $x5, x6$) dominated the distributions patterns of clay, sand, OM, Fe_o and, to a lesser extent, also those of Fe and K. By contrast, there was no systematic trend in the measurements in the orthogonal direction

NW – SE along the slope of the catchment. Furthermore, there were no evident relationships between the distribution patterns of the variables and topographical terrain attributes (not shown). The patterns suggested that the observations were spatially auto-correlated across several levels of the nested design, as the two measurements of a double very often were more similar to each other than to the values of the second double of the quadruple (see Figure 2 for grouping of locations to doubles, quadruples, etc.). Some of these doubles are marked in Figure 4 by dotted rectangles. Similarly, it was quite common that the observations of quadruples within octuples (short-dashed rectangles) and, less frequently, also of octuples within clusters (long-dashed rectangles) were more alike. For some variables (Fe, Al, Mn), also clusters appeared more homogeneous than cluster pairs (dot-dashed rectangles), suggesting that some of the variance components $\sigma^2_{0.6\mathrm{m}}$ to $\sigma^2_{20\mathrm{m}}$ were non-negligible.

Estimating the spatial scales of variation by REML

The REML fits confirmed that the means of the texture variables, of OM, $\mathrm{Fe_o}$ and Al differed significantly ($P < 0.05$) between the SW and NE parts of the catchment. For the other variables, the apparent differences visible in Figures 3 and 4 were not statistically significant.
Figure 5 shows for a selection of variables the estimated contribution of the various levels of the nested design (Figure 2 and Equation 1) to the total variance. Both the individual and the accumulated variance components (starting the summation from $\widehat{\sigma}^2_{0.2\mathrm{m}}$) are shown for the estimates computed by the function *lme* (constrained to positive values). For comparison, we also show the accumulated components for the unconstrained estimates $\widetilde{\sigma}^2$.
For most variables, the variance of the residual errors ($\widehat{\sigma}^2_{0.2\mathrm{m}}$) was the largest variance component: $\widehat{\sigma}^2_{0.2\mathrm{m}}$ generally amounted to 1/3 to 1/2 of the total variance, except for sand and pH (3/4 and 1/10 of the total variance, respectively). As suspected, some variance components associated with the larger spatial scales of the design were of the same order of magnitude as $\widehat{\sigma}^2_{0.2\mathrm{m}}$. There seemed to be three groups of variables with similar patterns of variation: for the textural variables, OM and $\mathrm{Fe_o}$, only the variance components associated with the scales 0.6 m and 2 m contributed substantially to the total variance in addition to $\widehat{\sigma}^2_{0.2\mathrm{m}}$. For most of the total, dithionite and oxalate extractable metal contents (some variables not shown) also $\widehat{\sigma}^2_{6\mathrm{m}}$ and/or $\widehat{\sigma}^2_{20\mathrm{m}}$ did not vanish. Finally, for pH, Ca and $\mathrm{P_{re}}$, also the variance component $\sigma^2_{>60\mathrm{m}}$ contributed noticeably to the total variance.
In many instances, the unconstrained and constrained estimates were identical. Negative variances were occasionally estimated for the scales $\geq$ 2 m, most often for >60 m, followed by 6 m. The covariance matrices, computed from the unconstrained estimates, were always positive definite, and the negative estimates were thus permissible. The unconstrained estimates fitted the data equally well or slightly better than the constrained *lme* estimates, but the differences in the maximized residual log-likelihood were small, except for the sand content. When negative variances were estimated then the estimates differed also for the adjacent levels of the design as observed by Webster *et al.* (2006). Nevertheless, the total variance of the observations hardly differed between unconstrained and constrained estimates.
The moduli of the negative estimates were mostly small. Figure 4 did not show any periodicity at the scale of 6 m (octuples) in the patterns of variation of the respective variables (OM, $\mathrm{Fe_o}$, Al, $\mathrm{P_{re}}$). Furthermore, the design had only four df for the scale >60 m (Table 1). Hence, we cannot safely conclude that there were regular features in the spatial distributions, apart from the SW/NE contrasts. Rather, we believe that the negative estimates signal zero contributions to the total variance at the respective levels of the nested design.
The uncertainty of some variance estimates was very large, as indicated by the partly unbounded confidence intervals, also shown in Figure 5. These confidence intervals were estimated from the Fisher Information matrix $\mathbf{J}(\ln \widehat{\sigma}^2_{0.2\mathrm{m}}, \ln \widehat{\sigma}^2_{0.6\mathrm{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\mathrm{m}})$. For OM, $\mathrm{Fe_d}$, Al and Mn, $\mathbf{J}$ was not positive definite, i.e. the log-likelihood surface did not have a well-defined maximum, and the covariance matrix of $\ln \widehat{\sigma}^2_{0.2\mathrm{m}}$ to $\ln \widehat{\sigma}^2_{>60\mathrm{m}}$ could not be computed from $\mathbf{J}(\ln \widehat{\sigma}^2_{0.2\mathrm{m}}, \ln \widehat{\sigma}^2_{0.6\mathrm{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\mathrm{m}})$. For the other variables $\mathbf{J}$ was positive definite, but some diagonal elements were very small, resulting in very large standard errors for the respective variance components. When we estimated the unconstrained variance components, the Hessian $\mathbf{J}(\widetilde{\sigma}^2_{0.2\mathrm{m}}, \widetilde{\sigma}^2_{0.6\mathrm{m}}, \ldots, \widetilde{\sigma}^2_{>60\mathrm{m}})$ was not positive definite for OM, K, $\mathrm{Fe_o}$, Al, Mn, $\mathrm{P_{re}}$, and for some of the remaining variables (clay, silt, $\mathrm{Fe_d}$), the standard errors of $\widetilde{\sigma}^2_{6\mathrm{m}}$, $\widetilde{\sigma}^2_{20\mathrm{m}}$and $\widetilde{\sigma}^2_{>60\mathrm{m}}$ were again large. This means that for all the variables except pH and Ca, the log-likelihood surface either did not have

a well-defined maximum or that the maximum occurred close to the boundary of the parameter space.
Figure 6 illustrates this for the clay and the Al content. For clay, the profile log-likelihood surface had a small peak at $\sigma^2_{6\mathrm{m}} \approx 10^{-5}$ along the abscissa, but there was no well defined maximum along the ordinate. For Al the maximum of the profile log-likelihood occurred very close to the origin. In both cases, the likelihood confidence regions had only well-defined upper bounds. The lower bounds coincided with the boundaries of the (constrained) parameter space. The weakly curved, straight shape of the contour lines implies that a (weighted) sum of the two variance components fitted the data roughly equally well. Thus, the data contained enough information to estimate the sum of the variances, but the information was insufficient to apportion it unambiguously to the two spatial scales.
To further explore the uncertainty of the variance estimates, we approximated 95% joint confidence regions for $\widehat{\sigma}^2_{0.2\mathrm{m}}, \ldots, \widehat{\sigma}^2_{>60\mathrm{m}}$ numerically. This is computationally quite demanding, and we did it therefore only for the Al and clay content, as these variables had posed identification problems, and for Ca and pH, which had well-defined peaks in their log-likelihood surfaces. Using a 20^5 grid to span the parameter space of the variance components $\sigma^2_{0.6\mathrm{m}}, \ldots, \sigma^2_{>60\mathrm{m}}$, we found between 593 847 (clay) and 1 077 864 (pH) 6-tuples on the boundary of the respective joint confidence regions. We accumulated the variances for all the 6-tuples, starting again with $\sigma^2_{0.2\mathrm{m}}$ and progressing with $\sigma^2_{0.6\mathrm{m}}$, etc., and computed for each level of the design the minimum and maximum of the accumulated variances. These extremes approximate a joint confidence region for the semivariances associated with the spatial scales of the nested design. The confidence regions are shown by the grey areas in Figure 7. Each panel shows, apart from the accumulated variance components of the constrained REML estimate (black curve), the accumulated components for six 6-tuples that 'touched' the upper bound of the confidence region at the six spatial scales (grey curves). We can see considerable variation in the shape of these curves. For the Al and clay content, a curve with a dominant nugget effect and curves implying an increase of the semivariance up to 0.6 m, 2 m, 6 m, 20 m or even over all the considered scales all lay within the confidence region and were thus compatible with the data at a 5% significance level. For pH and Ca, the shape of the curves varied less, especially at the short scales. Most curves suggested an increase of the semivariance up to >60 m. For all four variables, the width of the confidence regions increased with increasing spatial scale, in particular for pH and Ca. For these variables, the width spanned more than an order of magnitude at the scale >60 m, illustrating the large uncertainty of the estimated semivariance. However, unlike the confidence intervals derived from the asymptotic Gaussian distribution assumed for $\ln \widehat{\sigma}^2_{0.2\mathrm{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\mathrm{m}}$, the confidence regions were always bounded, even when some variance components could not be identified well (Al, clay).
As a means of mitigating ill-determined estimates, we fitted 'reduced' models (selected by AIC) that contained only a subset of the random effects of the full model (1). Table 2 lists the estimated variances, along with likelihood confidence intervals. For four variables, the confidence intervals are also shown in Figure 8, along with intervals computed from the Fisher Information of constrained and unconstrained estimates.
In general, the variances fitted for the 'reduced' models matched the respective estimates of the 'full' models fairly well. Likelihood ratio tests showed that the 'reduced' models fitted the data as well as the 'full' models and that they fitted the data significantly better than models with just the residual errors. Thus, all the variables, except the coarse sand content, were spatially auto-correlated, some even strongly.
For $\sigma^2_{0.2\mathrm{m}}$, the three types of confidence intervals matched quite well, but for the other variance components, there were some discrepancies: Intervals derived from $\mathbf{J}(\widetilde{\sigma}^2_{0.2\mathrm{m}}, \ldots \widetilde{\sigma}^2_{>60\mathrm{m}})$ were consistently the shortest, had the smallest lower and upper bounds, and were symmetrical about the REML estimates. The likelihood confidence intervals and intervals computed from $\mathbf{J}(\ln \widehat{\sigma}^2_{0.2\mathrm{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\mathrm{m}})$ were asymmetrical and, except for $\sigma^2_{20\mathrm{m}}$ (pH) and $\sigma^2_{>60\mathrm{m}}$(Ca), had similar width.

Precision of variance components estimated from balanced und unbalanced nested designs

Figures 5 and 7 suggest that in particular the variance components associated with the larger spatial scales were hard to estimate. Balanced designs always have a much smaller number of degrees of freedom at the larger than at the smaller spatial scales. Oliver & Webster (1986) and later Khattree

et al. (1997) remarked that this imbalance in the allocation of df either leads to unnecessary precision at the shorter spatial scales or to unreliable estimates at the larger scales, and therefore recommended unbalanced designs for variance component estimation. These designs allocate the degrees of freedom more evenly over all levels.

In order to see how much improvement might have been possible by using an unbalanced nested designs, we performed simulations in which we compared the balanced design of the survey (Figure 9, left panel) with three unbalanced staggered designs (Webster & Boag, 1992; Pettitt & McBratney, 1993; Khattree *et al.*, 1997). The first unbalanced design was staggered within cluster pairs (Figure 9, left panel) and involved only six observations per cluster pair. The second was staggered within octuples and had 16 observations per pair (solid black and dashed lines in Figure 9), and the last unbalanced design was staggered within quadruples with 24 observations per pair. For all designs, the sample size was 192, we used 32 (staggered within cluster pairs), 12 (staggered within octuples), eight (staggered within quadruples) and six (balanced) cluster pairs, respectively, to get the right sample size. Table 1 lists the degrees of freedoms associated with the six spatial scales in the four designs. The staggered-within-pairs design evenly distributed 32 df across all scales. The staggered-within-octuples had 48 df at the three shortest scales and still 10 df for >60 m. The staggered-within-quadruples design assigned the df more unevenly, only the distances 0.2 m and 0.6 m had the same df (64). Finally, the balanced design allocated many more df to the short than to the large distances. Since we estimated two fixed effects, just 4 df remained for the scale >60 m in the balanced design.

As set out above, we simulated 1000 data sets, estimated the variance components for each design and data set by REML, and characterized the statistical properties of the estimates by bias and root mean square error. The bias was always quite small: $\text{BIAS}^2/\text{RMSE}^2$ was at most 8% and often smaller (0 – 2%). Thus, the REML estimators were (practically) unbiased for all the designs, but they differed in their RMSEs (Figure 9, right panel). As to be expected, the variance components of the two smallest scales ($\sigma^2_{0.2\text{m}}$, $\sigma^2_{0.6\text{m}}$) were most precisely estimated by the balanced design, while the most precise estimates of $\sigma^2_{>60\text{m}}$ (largest scale) were obtained with the staggered-within-pairs design. The RMSE of $\sigma^2_{>60\text{m}}$ was more than twice as large for the balanced than for the staggered-within-pairs design. Conversely, the staggered-within-pairs design ranked last for $\sigma^2_{0.2\text{m}}$ to $\sigma^2_{6\text{m}}$. The two other staggered designs, which were less strongly unbalanced, were best at the intermediate scales ($\sigma^2_{0.6\text{m}}$ to $\sigma^2_{6\text{m}}$), with staggered-within-quadruples better at the short ($\sigma^2_{0.2\text{m}}$ to $\sigma^2_{6\text{m}}$) and staggered-within-octuples at the two largest scales ($\sigma^2_{20\text{m}}$, $\sigma^2_{>60\text{m}}$). If we compare the RMSEs with the magnitude of the variances used for simulating the data (the latter are shown by horizontal lines in Figure 9) the relative precision of the variance estimates, as expressed by RMSE/σ^2, deteriorated with decreasing σ^2. Irrespective of the design, the worst relative precision was found for $\sigma^2_{0.6\text{m}}$, $\sigma^2_{6\text{m}}$ and $\sigma^2_{20\text{m}}$. This shows that it is more difficult to estimate a small than a large variance component with the same relative precision.

The reversed ranking of the designs with respect to RMSE agreed only for $\sigma^2_{0.2\text{m}}$ and $\sigma^2_{>60\text{m}}$ with the allocated degrees of freedom (Table 1). For the other variance components, this was not the case. In particular, the staggered-within-pairs design gave the largest RMSE for $\sigma^2_{6\text{m}}$ and $\sigma^2_{20\text{m}}$, although its df were largest there. Thus, one cannot infer the precision of variance estimates at a given level of a nested design solely from the degrees of freedom allocated to this particular level. We suspect that the magnitude of the variances associated with the various levels of the design might have some influence.

As a further way to compare the merits of the designs, we counted the number of simulated data sets for which the moduli of the diagonal elements of $\mathbf{J}(\ln \widehat{\sigma}^2_{0.2\text{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\text{m}})$ were $< 10^{-4}$. The log-likelihood surface was then flat with respect to $\ln \widehat{\sigma}^2_{0.2\text{m}}, \ldots, \ln \widehat{\sigma}^2_{>60\text{m}}$. This happened in particular when the maximum of the likelihood lay very close to the boundary of the parameter space. Table 3 lists the number of times this occurred for the various variance components. Irrespective of the design, $\sigma^2_{0.2\text{m}}$ was always unambiguously estimated. The balanced design was best for estimating $\sigma^2_{0.6\text{m}}$ and $\sigma^2_{2\text{m}}$, but worst for the other components, especially for $\sigma^2_{6\text{m}}$ and $\sigma^2_{>60\text{m}}$. The staggered-within-pairs design was best for $\sigma^2_{>60\text{m}}$, but worst for $\sigma^2_{2\text{m}}$. The staggered-within-octuples design was best for $\sigma^2_{6\text{m}}$ and $\sigma^2_{20\text{m}}$ and worst for $\sigma^2_{0.6\text{m}}$, but performed quite well at the other scales. On average, this design was best. The staggered-within-quadruples design did not perform much better than the balanced design which produced on average the largest number of ill-determined estimates. The largest percentages of ill-determined estimates were found for the small variance components ($\sigma^2_{0.6\text{m}}$, $\sigma^2_{6\text{m}}$, $\sigma^2_{20\text{m}}$), which is

in agreement with their large relative RMSEs.
On the whole, taking both the RMSE and the tendency to yield ill-determined estimates as criteria, the staggered-within-octuples design performed best. This suggests that neither a completely balanced, nor a very strongly unbalanced design may be optimal. This result contrasts the practice of nested sampling in soil science where either balanced or very strongly unbalanced designs seem to have been favoured (see examples in Webster *et al.*, 2006; Lark, 2005; Corstanje *et al.*, 2007; Lark & Corstanje, 2009). However, the evidence presented here relates to just one simulation scenario, and more work is required before wider generalizations can be made.

Relevance of the results for the 'Chicken Creek' ecosystem research project

All response variables, except for the coarse sand content, were, apart from a possible SW/NE contrast, spatially auto-correlated. For coarse sand, $\sigma^2_{0.2\mathrm{m}}$ was by far the largest variance component, and the other variances were jointly not significantly different from zero ($p = 0.04$). But for all the other variables, likelihood ratio tests clearly refuted models that pre-supposed spatial independence and assumed that all variance components except $\sigma^2_{0.2\mathrm{m}}$ were equal to zero. However, it was not possible to say which components were actually non-vanishing. Soil pH and Ca were the only variables for which we could estimate all variance components with reasonable precision. Even for these variables, the uncertainty of the estimated semivariances remained considerable (Figure 7). For the other variables, the log-likelihood did not show a well-defined maximum with respect to the logarithm of the variance components. Thus, the 192 available observations in general lacked the necessary information to apportion the total variance to the various spatial scales. We therefore cannot soundly relate the results of our analysis to the spatial information that we have about the construction of the 'Chicken Creek' catchment.
Our simulations indicate that somewhat more reliable estimates could have been obtained, had we used a staggered unbalanced design, but very likely, the gain would have been limited. Thus, a marked improvement could have been achieved only by increasing the sample size substantially. To see what sample size would be required, we ran the simulations with larger data sets and estimated the variance components with the balanced and staggered-within-octuples design from 512 and and 1024 observations, respectively. The average percentage of simulated data sets with ill-determined estimates decreased to 3.9% (1.3%) for the balanced and 3.3% (1.8%) for the staggered design for estimates computed from 512 (1024) observations. Thus, it appears that far more observations would have been required to estimate the spatial variance components reliably. The sample size and the number of levels of our survey are not unusual in comparison to other nested sampling studies published in the literature. As examples, for a design with five levels Oliver & Webster (1987) used 108 observations, and Webster & Boag (1992) chose 107 and (Oliver & Badr, 1995) 105 observations for seven levels. Cole (2009) was dissatisfied with the precision of variance component estimates obtained by a staggered design and pointed out that confidence intervals of variance components were rarely reported in ecological studies. This might explain why the apparently poor precision of variance component estimates in nested sampling surveys have passed largely unnoticed so far.
Our results suggest that the development of the ecosystem in the 'Chicken Creek' catchment did not start from a horizontally unstructured soil. For most variables, we found substantial spatial variation, the coefficient of variations ranged from about 15% for the sand to more than 70% for the Ca content, and strong auto-correlation three years after the construction of the catchment. While the spatial distribution patterns of some soil characteristics may have changed during the first three years after construction of the catchment, for example the organic matter content, soluble P, etc., substantial change in spatial patterns is rather unlikely for others, (such as total metal contents). It is therefore likely that the spatial structures are primarily related to the initial heterogeneity resulting from the construction process.

Conclusions

Our study shows how standard likelihood inference tools (Fisher Information, profile likelihood, likelihood confidence regions) can be used to assess and estimate the uncertainty of spatial variance components in nested sampling. To our knowledge, this has not been done before in an application of nested sampling in soil science.
This is rather surprising, because our case study shows that it may be difficult or even impossible to

identify variance components with acceptable precision. We showed that the joint confidence regions for the semivariances associated with the levels of the nested design were wide and did not allow us to identify unambiguously the spatial scales over which the observations were correlated. This is unfortunate because we wanted to characterize the scales of spatial variation of soil properties at the onset of ecosystem development in the artificially created 'Chicken Creek' catchment, and we hoped that our study would provide a sound basis for detecting changes in the auto-correlation patterns later on. Compared with other nested sampling studies, the sample size (192) and the number of nested levels (six) of our survey were not untypical. Using an unbalanced survey design would have provided somewhat more precise estimates of variance components at the larger scales, but estimates with acceptable precision would have required a much larger sample size. We suspect that the difficulties that we encountered are a general problem of nested sampling. Thus, we see a need for a comprehensive assessment of the merits and disadvantages of the nested sampling methodology. The following questions should be addressed:

i) What sample size is required to estimate a given number of nested variance components with sufficient precision?

ii) What degree of imbalance of a design is optimal?

iii) Does the choice of an optimal design depend on the number of levels and on the relative size of the variances associated with the various levels?

iv) How well do confidence intervals derived from the asymptotic normal distribution of the REML estimates characterize the estimation uncertainty? Should likelihood-based confidence regions be preferred?

The methods proposed here may provide a basis to tackle some of these questions.

Acknowledgements

This study is part of the Transregional Collaborative Research Centre 38 (SFB/TRR 38) which is financially supported by the Deutsche Forschungsgemeinschaft (DFG, Bonn) and the Brandenburg Ministry of Science, Research and Culture (MWFK, Potsdam). The authors thank Vattenfall Europe Mining AG for providing the research site and R. Spröte, P. Lange, S. Chabbi and B. Wulfert for valuable support during the field work. We thank an anonymous referee for useful comments on an earlier version of the manuscript.

References

Cole, R.G. 2009. Staggered nested designs to assess scales of variability: The advantages of a spatially explicit analysis. *Environmental Monitoring & Assessment*, **153**, 427–434.

Corstanje, R., Schulin, R. & Lark, R.M. 2007. Scale-dependent relationships between soil organic carbon and urease activity. *European Journal of Soil Science*, **58**, 1087–1095.

Gerwin, W., Schaaf, W., Biemelt, D., Fischer, A., Winter, S. & Hüttl, R.F. 2009. The artificial catchment 'Chicken Creek' (Lusatia, Germany) — A landscape laboratory for interdisciplinary studies of initial ecosystem development. *Ecological Engineering*, **35**, 1786–1796.

Jiang, J. 2007. *Linear and Generalized Linear Mixed Models and Their Applications*. Springer, New York.

Kendzia, G., Reissmann, R. & Neumann, T. 2008. Targeted development of wetland habitats for nature conservation fed by natural inflow in the post-mining landscape of Lusatia. *World of Mining—Surface & Underground*, **60**, 88–95.

Khattree, R., Naik, D.N. & Mason, R.L. 1997. Estimation of variance components in staggered nested designs. *Journal of Applied Statistics*, **24**, 395–408.

Lark, R.M. 2005. Exploring scale-dependent correlation of soil properties by nested sampling. *European Journal of Soil Science*, **56**, 307–317.

Lark, R.M. & Corstanje, R. 2009. Non-homogeneity of variance components from spatially nested sampling of the soil. *European Journal of Soil Science*, **60**, 443–452.

Maronna, R.A., Martin, R.D. & Yohai, V.J. 2006. *Robust Statistics Theory and Methods*. John Wiley & Sons, Chichester.

Mehra, O.P. & Jackson, M.L. 1960. Iron oxide removal from soils and clays by a dithionite-citrate system buffered with sodium bicarbonate. In: *Proceedings of the 7th National Conference on Clays and Clay Minerals* (ed. A. Swineford), pp. 317–327, Pergamon Press, London.

Miesch, A.T. 1975. Variograms and variance components in geochemistry and ore evaluation. *Geological Society of America. Memoir*, **142**, 333–340.

Oliver, M.A. & Badr, I. 1995. Determining the spatial scale of variation in soil radon concentration. *Mathematical Geology*, **27**, 893–922.

Oliver, M.A. & Webster, R.A. 1986. Combining nested and linear sampling for determining the scale and form of spatial variation of regionalized variables. *Geographical Analysis*, **18**, 227–242.

Oliver, M.A. & Webster, R. 1987. The elucidation of soil pattern in the Wyre Forest of the West Midlands, England. II. Spatial distribution. *Journal of Soil Science*, **38**, 293–307.

Pettitt, A.N. & McBratney, A.B. 1993. Sampling designs for estimating spatial variance components. *Applied Statistics*, **42**, 185–209.

Pinheiro, J.C. & Bates, D.M. 2000. *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

R Development Core Team 2009. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna.

Rawlins, B.G., Scheib, A.J., Lark, R.M. & Lister, T.R. 2009. Sampling and analytical plus subsampling variance components for five soil indicators observed at regional scale. *European Journal of Soil Science*, **60**, 740–747.

Rousseeuw, P.J. & Croux, C. 1993. Alternatives to median absolute deviation. *Journal of the American Statistical Association*, **88**, 1273–1283.

Saggar, S., Hedley, M.J. & White, R. 1990. A simplified resin membrane technique for extracting phosphorus from soil. *Fertilizer Research*, **24**, 173–180.

Schwertmann, U. 1964. Differenzierung der Eisenoxide des Bodens durch Extraktion mit Ammoniumoxalat. *Zeitschrift für Pflanzenernährung, Düngung und Bodenkunde*, **105**, 194–202.

Stram, D.O. & Lee, J.W. 1994. Variance components testing in the longitudinal mixed-effects models. *Biometrics*, **50**, 1171–1177.

Turner, M.G. 2005. Landscape ecology: What is the state of the science? *Annual Review of Ecology, Evolution & Systematics*, **36**, 319–344.

Van Veldhoven, P.P. & Mannaerts, G.P. 1987. Inorganic and organic phosphate measurements in the nanomolar range. *Analytical Biochemistry*, **161**, 45–48.

Webster, R. & Boag, B. 1992. Geostatistical analysis of cyst nematodes in soil. *Journal of Soil Science*, **43**, 583–595.

Webster, R. & Oliver, M.A. 2007. *Geostatistics for Environmental Scientists*. John Wiley & Sons, New York, second edition.

Webster, R., Welham, S.J., Potts, J.M. & Oliver, M.A. 2006. Estimating the spatial scales of regionalized variables by nested sampling, hierarchical analysis of variance and residual maximum likelihood. *Computers & Geosciences*, **32**, 1320–1333.

Appendix

To compute a joint confidence region for a set of q variance components we chose $q' = q - 1$ variance components, say $\sigma^2_{i_k}$, $k = 1, 2, \ldots, q'$, and placed a grid into a hyper-rectangle in the q'-dimensional parameter space spanned by the $\sigma^2_{i_k}$. For each $\sigma^2_{i_k}$, $k = 1, 2, \ldots, q'$, the values of the grid nodes were bounded by the roots of:

$$2[L_{\max} - L_{\mathrm{p}}(\sigma^2_{i_k})] = \chi^2_q(1 - \alpha). \tag{A1}$$

Occasionally, the above equation had no lower root. When this happened we set the lower bound of $\sigma^2_{i_k}$ to $\widehat{\sigma}^2_{i_k}/10^m$ with $m \geq 3$ (the value of m depended on $\partial L_{\mathrm{p}}/\partial\sigma^2_{i_k}$). The nodes of the grid were equally spaced between these bounds on the scale of the standard deviations. Next we checked whether the inequality:

$$2[L_{\max} - L_{\mathrm{p}}(\sigma^2_{i_1}, \ldots, \sigma^2_{i_{q'}})] < \chi^2_q(1 - \alpha) \tag{A2}$$

held. If this was not the case then we moved to the next node, because any q-tuple, irrespective of the chosen value for the qth variance component, say σ^2_q, lies then *outside* of the joint confidence region. If the inequality of Equation A2 was satisfied then we computed for the values $\sigma^2_{i_1}, \ldots, \sigma^2_{i_{q'}}$ of the current node the roots of:

$$2[L_{\max} - L_{\mathrm{p}}(\sigma^2_q; \sigma^2_{i_1}, \ldots, \sigma^2_{i_{q'}})] = \chi^2_q(1 - \alpha) \tag{A3}$$

with respect to σ^2_q. The resulting q-tuples $\sigma^2_q, \sigma^2_{i_1}, \ldots, \sigma^2_{i_{q'}}$ lie on the boundary of the joint confidence region. We implemented this algorithm in an *R* function, which is available upon request from the corresponding author.

Figure 1: Areal view of the artificial catchment 'Chicken Creek' during construction in April 2005. The points mark locations chosen in 2008 for nested sampling (position of quadruples). The solid line is the final boundary of the catchment. The insets show enlarged views of sections A, B and C (mesh width of insets: 5 m). Source: Vattenfall Europe Mining AG.

Figure 2: Position of the 12 clusters of sampled locations in 'Chicken Creek' catchment (left) and schematic view of the arrangement of the 16 locations within a cluster (right). The grey level in the left panel codes the elevation, measured in April 2009 by means of a laser scan. The shaded rectangles of the right panel illustrate the grouping of locations to clusters (C_{ij}), octuples (O_{ijk}), quadruples ($Q_{ij\ldots l}$), and doubles ($D_{ij\ldots m}$). Source of elevation data: Vattenfall Mining Europe AG.

Figure 3: Histograms of selected variables and probability density estimates of observations collected in the SW (dashed) and NE (solid line) parts of the catchment. The density estimates were multiplied by 0.5 so that the areas under both curves sum to one.

Figure 4: Spatial distribution of the measurements, c, of selected variables within the 12 clusters (D2 – N6) of the nested sampling design (disc area $\propto (c - \min(c))/(\max(c) - \min(c))$, coordinates of sampled locations rotated and shifted).

Figure 5: Variance components estimated by *R* function *lme* (•, estimates constrained to be positive) with approximate 95% confidence intervals (vertical grey lines), plotted against the spatial scales of the nested design. The accumulated variance components are shown by the black solid lines (*lme* estimates) and the dashed grey lines (unconstrained estimates). The maximized residual log-likelihood of the constrained and unconstrained estimates are listed in the top right hand part of each panel (the upper number refers to the constrained, the lower to the unconstrained estimates). For some variables, the confidence intervals could not be computed (see text for an explanation).

Figure 6: Contour plots of the profile log-likelihood as a function of two variance components, for the total Al and clay content (• REML estimate; thick grey line: upper bound of joint 95% confidence region).

Figure 7: Approximate 95% log-likelihood confidence regions for the accumulated variance components (semivariances) plotted against the spatial scales of the nested design for four variables. The solid black lines are the REML estimates, shown also in Figure 5, the grey lines show semivariance curves that hit the upper bounds of the confidence regions at the six spatial scales. For Al clay a curve with a dominant nugget effect (solid grey line) and curves implying an increase of the semivariance up to 0.6 m (short-dashed), 2 m (dotted), 6 m (dot-dashed), 20 m (long-dashed) or even over all the considered scales (dot-dot-dashed) all lay within the confidence region.

Figure 8: 95% confidence intervals for the variance components of the 'reduced' models computed from the Fisher Information matrices of the unconstrained (grey lines with black dots), the constrained estimates (grey lines) and based on likelihood ratio tests (black lines; • REML estimates).

Figure 9: Schematic representation of the (un-)balanced nested sampling designs for a cluster pair (left) used in the simulations to assess the precision of variance component estimates (grey lines: balanced design; solid black lines: staggered-within-pairs design; solid black and dashed black lines: staggered-within-octuples design; solid black, dashed and dotted lines: staggered-within-quadruples design) and root mean square errors of variance component estimates (right; horizontal lines: variances used to simulate the random effects).

Table 1: Number of degrees of freedom at the various levels of the nested designs used in the simulations to assess the precision of variance component estimates (a further two degrees of freedom were associated with the fixed effects).

Design	$\sigma^2_{0.2\mathrm{m}}$	$\sigma^2_{0.6\mathrm{m}}$	$\sigma^2_{2\mathrm{m}}$	$\sigma^2_{6\mathrm{m}}$	$\sigma^2_{20\mathrm{m}}$	$\sigma^2_{>60\mathrm{m}}$
Balanced	96	48	24	12	6	4
Staggered within quadruples	64	64	32	16	8	6
Staggered within octuples	48	48	48	24	12	10
Staggered within cluster pairs	32	32	32	32	32	30

Table 2: Estimates of variance components $\times 10^4$ of models that minimized the Akaike Information criterion (numbers in parentheses: approximate 95% confidence interval based likelihood ratio tests; empty cells: components not part of the reduced model; $\Delta L, \mathrm{df}, p$: test statistic, degrees of freedom and p-value of likelihood ratio test comparing the reduced with the full model [Eq. 1]).

	$\widehat{\sigma}^2_{0.2\mathrm{m}}$	$\widehat{\sigma}^2_{0.6\mathrm{m}}$	$\widehat{\sigma}^2_{2\mathrm{m}}$	$\widehat{\sigma}^2_{6\mathrm{m}}$	$\widehat{\sigma}^2_{20\mathrm{m}}$	$\widehat{\sigma}^2_{>60\mathrm{m}}$	ΔL	df	p
Clay	2.01	1.11	1.56				1.48	3	0.69
	(1.54, 2.71)	(0.36, 2.28)	(0.56, 3.04)						
Sand	30.0		8.53				2.05	4	0.73
	(24.0, 38.1)		(3.0, 17.5)						
pH	57.7	22.3	92.9		343		3.26	2	0.20
	(44.0, 77.6)	(3.4, 50.6)	(50, 170)		(143, 1010)				
OM	0.223	0.102	0.0923				1.72	3	0.63
	(0.171, 0.300)	(0.025, 0.220)	(0.007, 0.207)						
K	53.0	21.9		25.4			2.08	3	0.56
	(40.3, 71.7)	(5.9, 43.2)		(8.8, 60.6)					
Ca	708		403	285		890	1.47	2	0.48
	(566, 903)		(403, 908)	(16, 924)		(336, 3020)			
Fe	159	94.3		42.4	62.9		0.65	2	0.72
	(121, 215)	(40, 170)		(0, 175)	(0, 254)				
$\mathrm{Fe_o}$	174	130	271				1.13	3	0.77
	(133, 235)	(56, 248)	(138, 484)						
$\mathrm{Fe_d}$	177	138		61.9	118		0.00	2	1.00
	(134, 239)	(69, 236)		(0, 242)	(0, 438)				
Al	99.7	66.2			27.5		1.43	3	0.70
	(76, 135)	(32, 112)			(4, 103)				
Mn	119	52.2		37.9	75.7		0.01	2	1.00
	(91, 161)	(16, 101)		(1, 142)	(0, 272)				
$\mathrm{P_{re}}$	365		287			343	1.45	3	0.70
	(292, 464)		(160, 508)			(77, 2300)			

Table 3: Percentage of simulated data sets for which the diagonal elements of the Fisher Information were very small (average: mean percentage for $\sigma^2_{0.6\mathrm{m}}$ to $\sigma^2_{>60\mathrm{m}}$).

Design	$\sigma^2_{0.2\mathrm{m}}$	$\sigma^2_{0.6\mathrm{m}}$	$\sigma^2_{2\mathrm{m}}$	$\sigma^2_{6\mathrm{m}}$	$\sigma^2_{20\mathrm{m}}$	$\sigma^2_{>60\mathrm{m}}$	Average
Balanced	0	10.9	0.0	21.9	23.1	7.9	12.8
Staggered within quadruples	0	15.5	0.2	19.0	23.2	2.5	12.1
Staggered within octuples	0	16.8	1.0	14.4	18.4	0.7	10.3
Staggered within cluster pairs	0	16.6	2.7	17.4	19.3	0.2	11.2

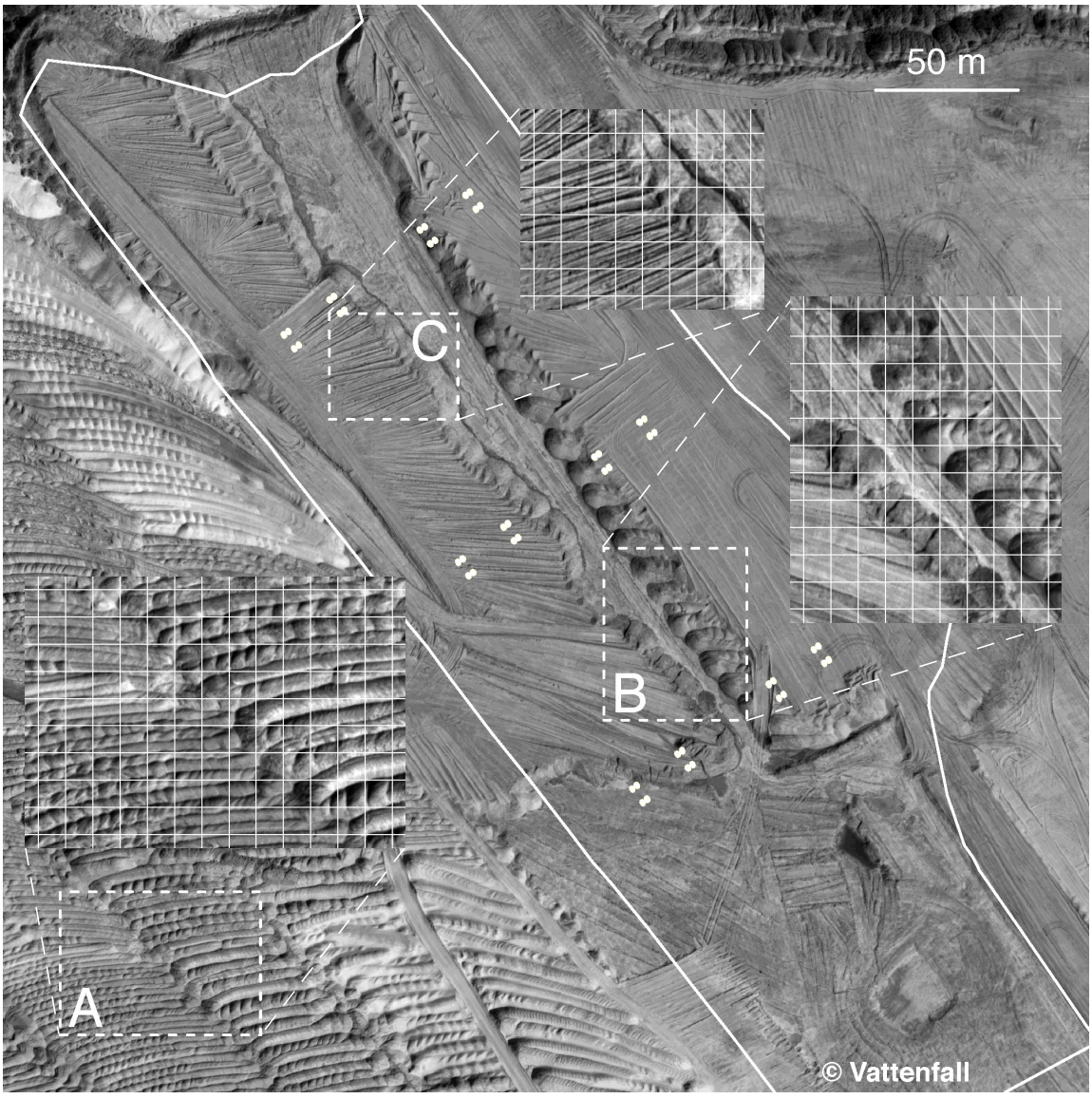

Figure 1:

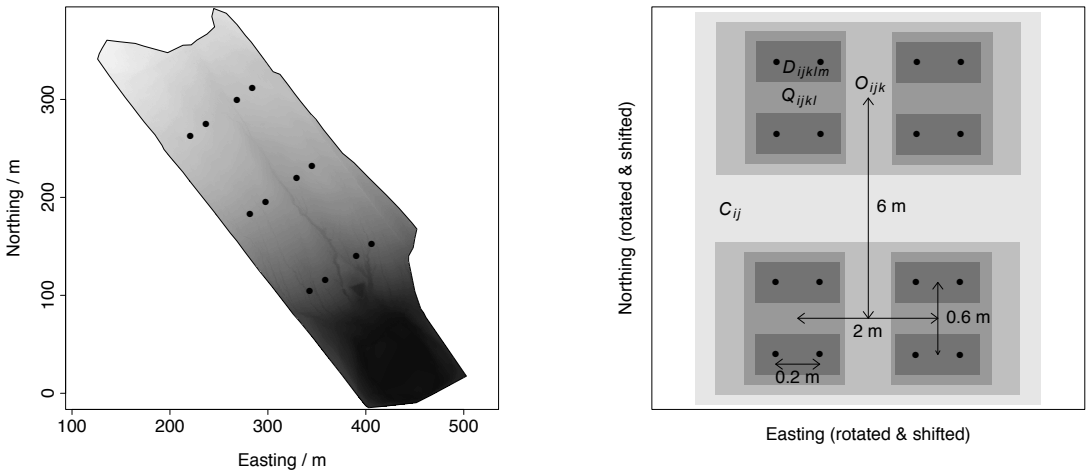
Northing / m
0
100
200
300
100
200
300
400
500
Easting / m
Northing (rotated & shifted)
D_{ijklm}
Q_{ijkl}
O_{ijk}
C_{ij}
6 m
2 m
0.6 m
0.2 m
Easting (rotated & shifted)

Figure 2:

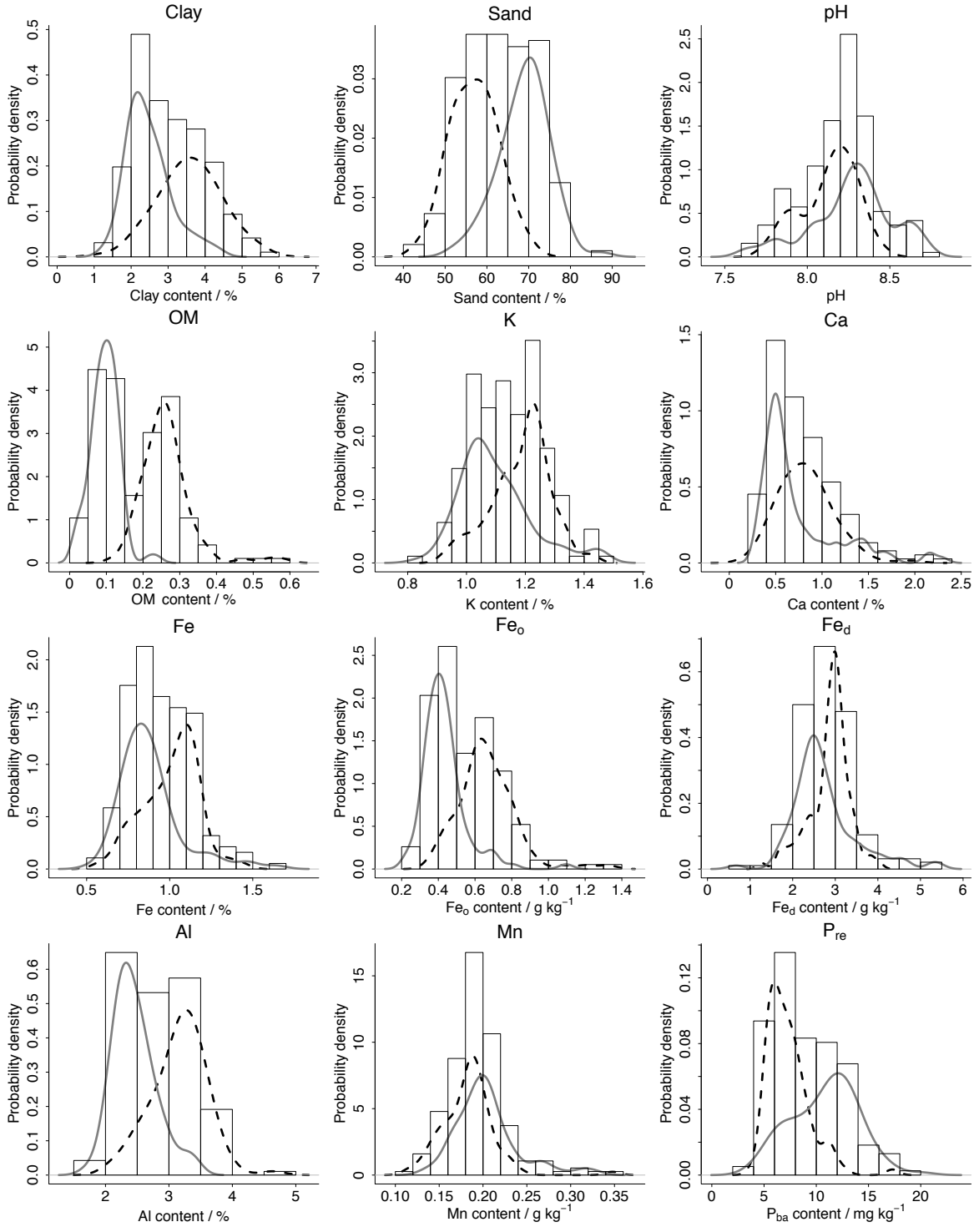
Clay
Probability density
Clay content / %
Sand
Sand content / %
pH
pH
OM
OM content / %
K
K content / %
Ca
Ca content / %
Fe
Fe content / %
Fe$_o$
Fe$_o$ content / g kg^{-1}
Fe$_d$
Fe$_d$ content / g kg^{-1}
Al
Al content / %
Mn
Mn content / g kg^{-1}
P$_{re}$
P$_{ba}$ content / mg kg^{-1}

Figure 3:

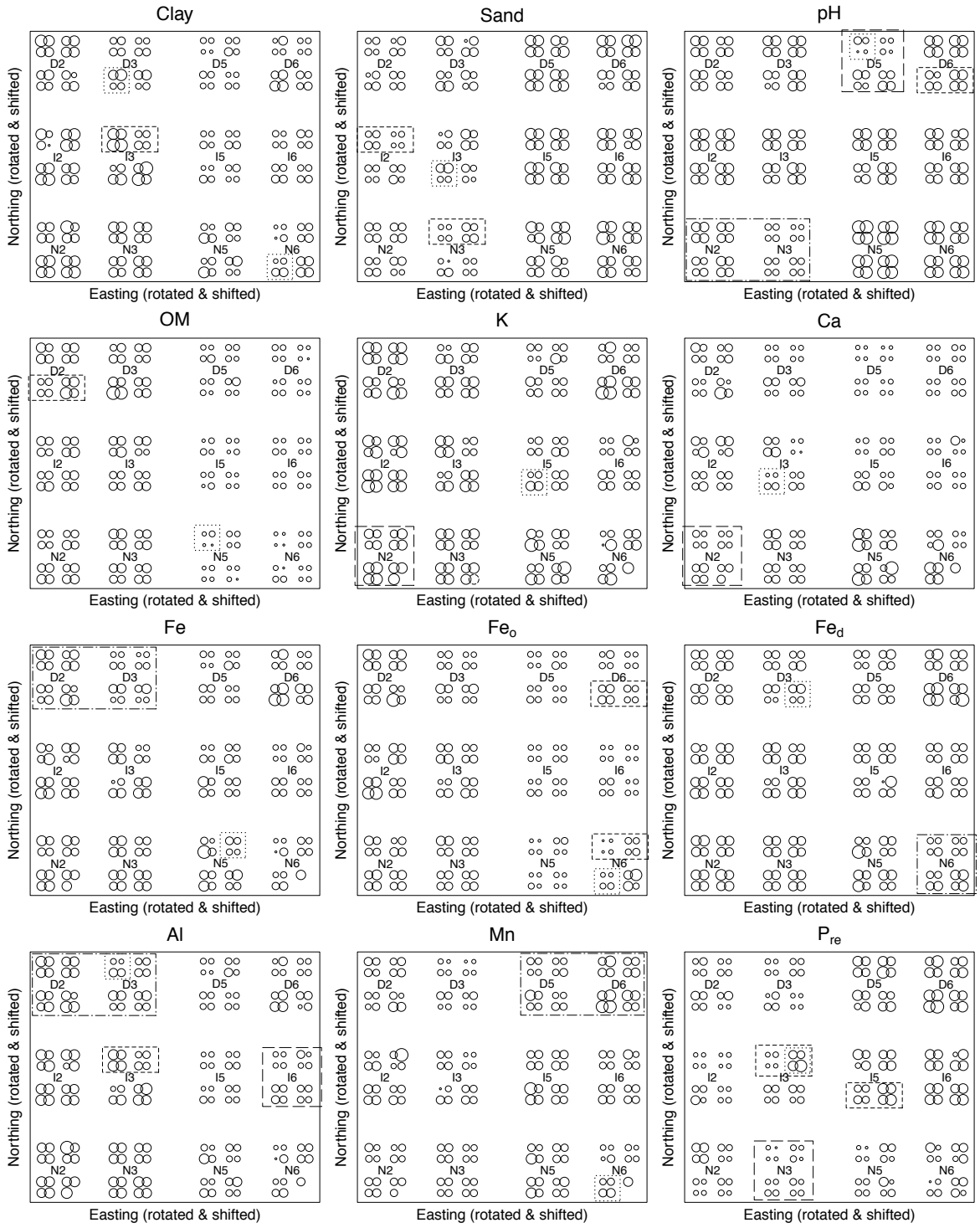
Clay
Sand
pH
OM
K
Ca
Fe
Fe$_o$
Fe$_d$
Al
Mn
P$_{re}$
Northing (rotated & shifted)
Easting (rotated & shifted)
D2
D3
D5
D6
I2
I3
I5
I6
N2
N3
N5
N6

Figure 4:

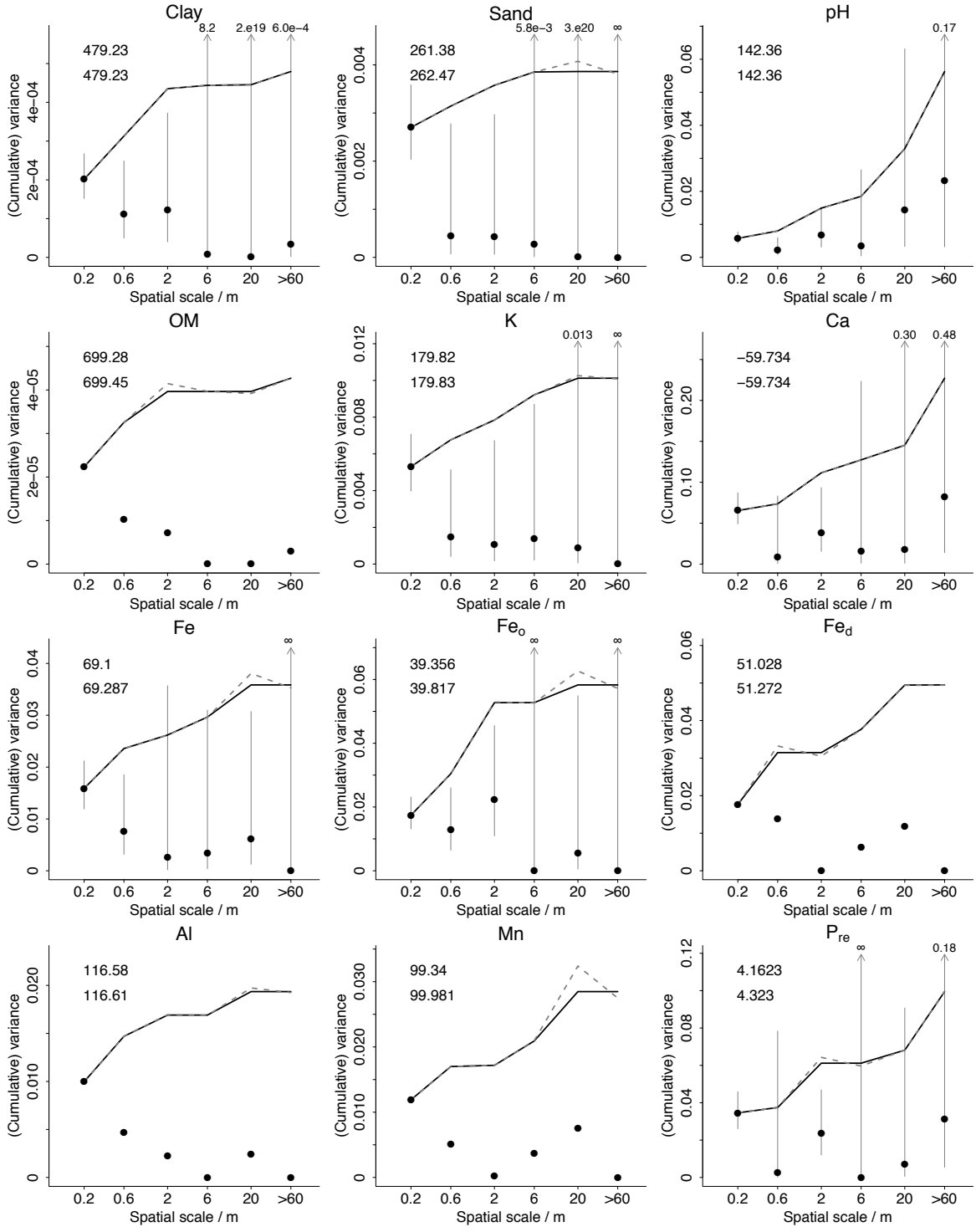
Clay
8.2
2.e19
6.0e−4
479.23
479.23
Sand
5.8e−3
3.e20
∞
261.38
262.47
pH
0.17
142.36
142.36
OM
699.28
699.45
K
0.013
∞
179.82
179.83
Ca
0.30
0.48
−59.734
−59.734
Fe
∞
69.1
69.287
Fe$_o$
∞
∞
39.356
39.817
Fe$_d$
51.028
51.272
Al
116.58
116.61
Mn
99.34
99.981
P$_{re}$
∞
0.18
4.1623
4.323
(Cumulative) variance
Spatial scale / m
0.2
0.6
2
6
20
>60

Figure 5:

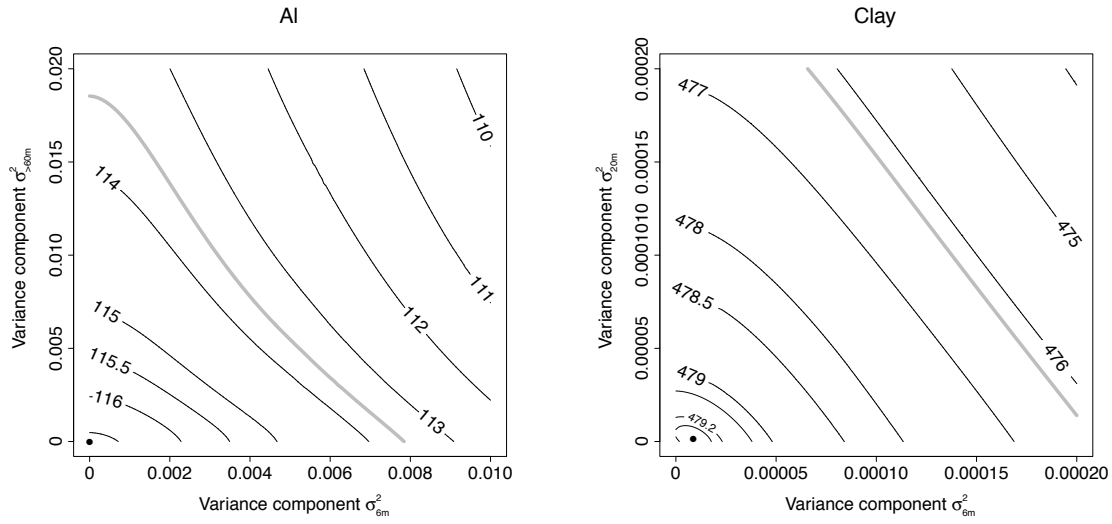
AI
Variance component $\sigma^2_{>60m}$
0
0.005
0.010
0.015
0.020
0
0.002
0.004
0.006
0.008
0.010
Variance component σ^2_{6m}
110
111
112
113
114
115
115.5
116
Clay
Variance component σ^2_{20m}
0
0.00005
0.0001010
0.00015
0.00020
0
0.00005
0.00010
0.00015
0.00020
Variance component σ^2_{6m}
475
476
477
478
478.5
479
479.2

Figure 6:

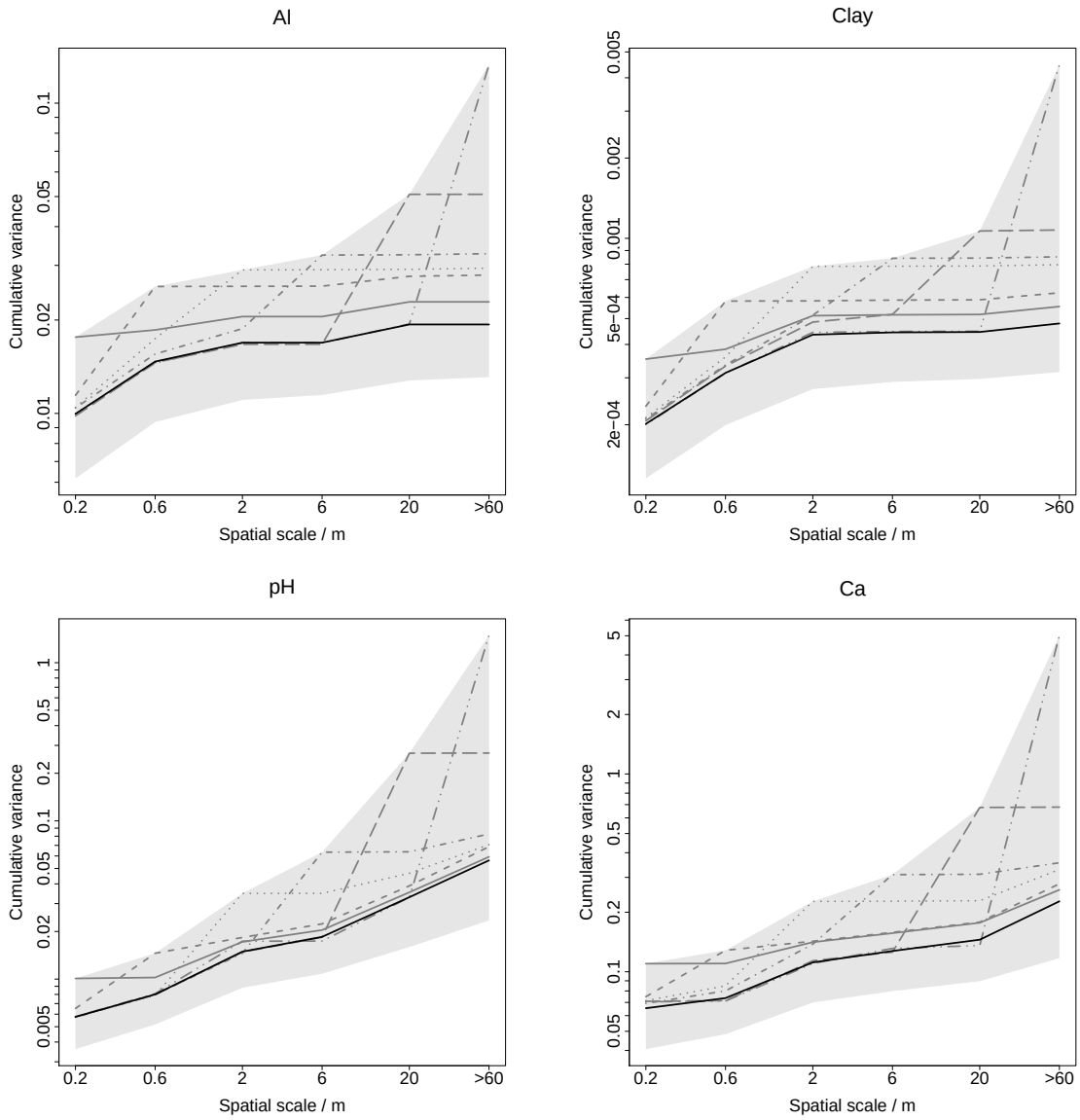
Al
Clay
pH
Ca
Cumulative variance
Spatial scale / m
0.2
0.6
2
6
20
>60
0.01
0.02
0.05
0.1
2e−04
5e−04
0.001
0.002
0.005
0.005
0.02
0.05
0.1
0.2
0.5
1
0.05
0.1
0.2
0.5
1
2
5

Figure 7:

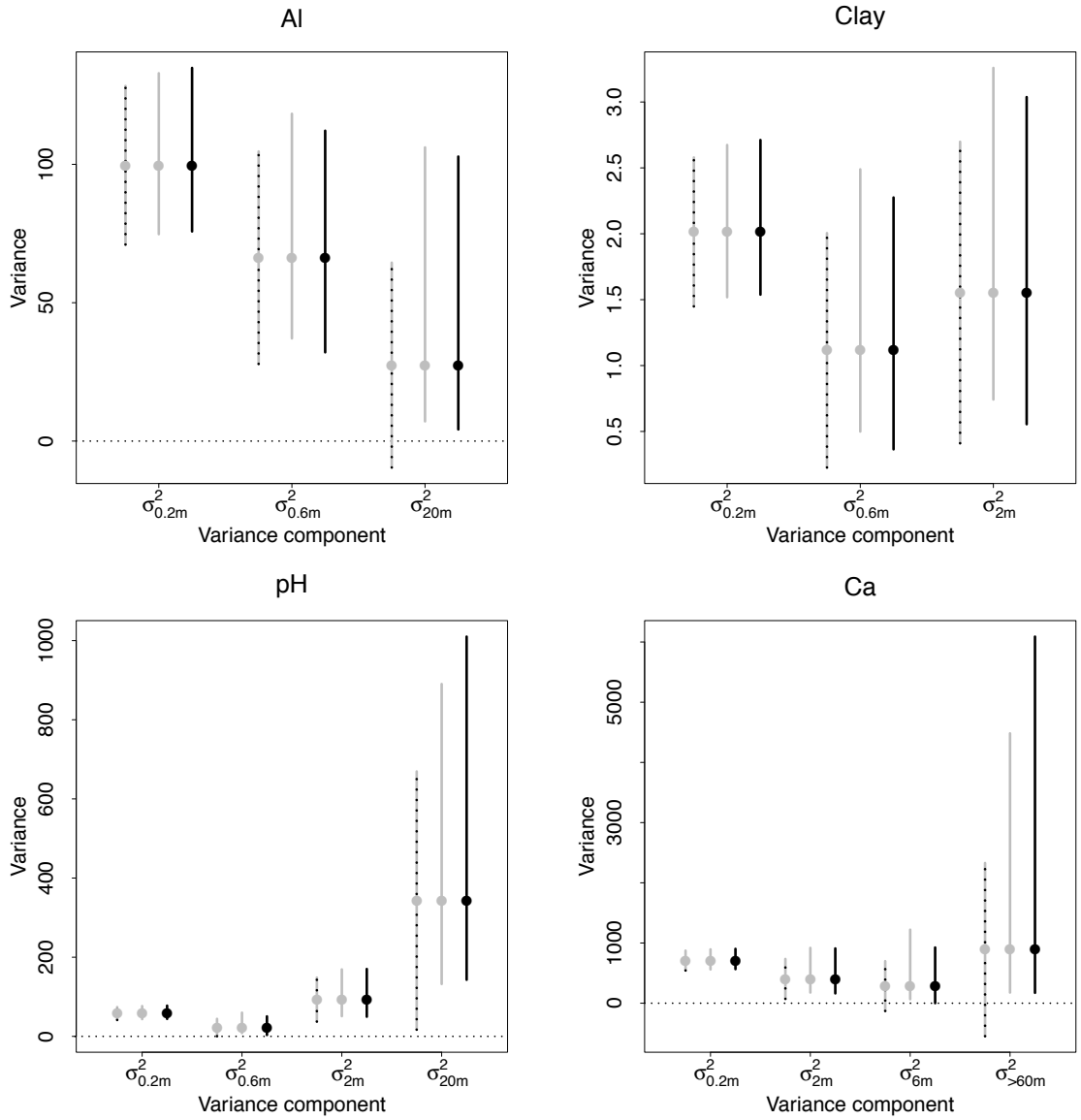
Al
Variance
0
50
100
$\sigma^2_{0.2m}$
$\sigma^2_{0.6m}$
σ^2_{20m}
Variance component
Clay
Variance
0.5
1.0
1.5
2.0
2.5
3.0
$\sigma^2_{0.2m}$
$\sigma^2_{0.6m}$
σ^2_{2m}
Variance component
pH
Variance
0
200
400
600
800
1000
$\sigma^2_{0.2m}$
$\sigma^2_{0.6m}$
σ^2_{2m}
σ^2_{20m}
Variance component
Ca
Variance
0
1000
3000
5000
$\sigma^2_{0.2m}$
σ^2_{2m}
σ^2_{6m}
$\sigma^2_{>60m}$
Variance component

Figure 8:

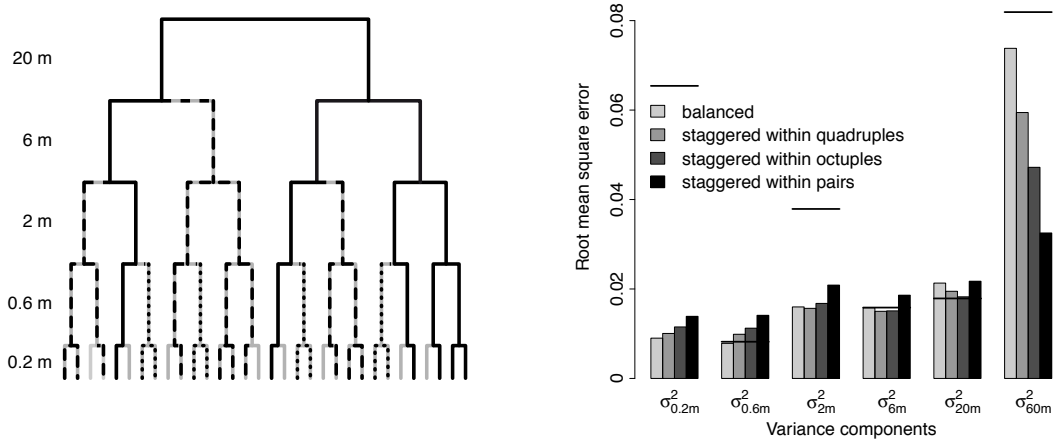
20 m
6 m
2 m
0.6 m
0.2 m
Root mean square error
0
0.02
0.04
0.06
0.08
balanced
staggered within quadruples
staggered within octuples
staggered within pairs
$\sigma^2_{0.2m}$
$\sigma^2_{0.6m}$
σ^2_{2m}
σ^2_{6m}
σ^2_{20m}
σ^2_{60m}
Variance components

Figure 9: