Distinguishing spatially correlated random variation in soil from a 'pure nugget' process

R.M. Lark[1]

British Geological Survey, Keyworth, Nottinghamshire NG12 5GG, U.K.

Abstract

In most spatial analysis of soil variation it is assumed that the random variation not captured by fixed effects (class means or continuous covariates) is spatially dependent. It is proposed that this should be tested formally, both to justify the kriging component in subsequent spatial prediction and as evidence of the extent to which the included fixed effects have succeeded in accounting for soil variation that is spatially dependent at the scales resolved by the soil sampling. A formal test is possible by computing the log ratio of the likelihoods for a full model with spatially dependent random effects and a null model which is pure nugget. It is shown that the sampling distribution of the log likelihood-ratio under the null model is not $\chi^2(p)$ where p is the number of additional random effects parameters in the model with spatial dependence. This is because, while the null model is nested in the full model, parameters of the full model take bounding values in the null case. The sampling distribution may be computed by Monte Carlo simulations. It is shown that the power to reject the null model by the log likelihood-ratio test depends on the importance of the nugget effect in the underlying model, and on the sampling scheme. In many circumstances it may be hard to demonstrate spatial dependence. The recommended procedure was applied to some data on the organic carbon content of the topsoil and subsoil of a field in England. This was modelled either with the overall mean the only fixed effects, or with separate means for different soil map units as fixed effects. There was significant evidence for spatial dependence in the random effects at both depths when the overall mean was the only fixed effect. When map unit means were used as fixed effects there was significant, though weaker, spatial dependence in the topsoil, but the null model could not be rejected for the subsoil. This has implications for any further sampling to map organic carbon in the subsoil.

Keywords: Linear mixed model; Nugget; Spatial dependence; Log likelihood-ratio.

[1] Corresponding author: *E-mail address*: mlark@nerc.ac.uk (R.M. Lark).

1. Introduction

In early statistical studies on soil variability and prediction from soil surveys (Webster and Beckett, 1968) a simple statistical model was used, implicitly or explicitly. Under this model the value of a soil property at a set of n locations, $\mathbf{S}$, is a random variate, $\mathbf{Y}(\mathbf{S})$, where

$$\mathbf{Y}(\mathbf{S}) \ = \ \mathbf{X}\boldsymbol{\tau} \ + \ \boldsymbol{\varepsilon}, \tag{1}$$

$\mathbf{X}$ is an $n \times p$ design matrix which associates each location in the set with one of p soil map units, $\boldsymbol{\tau}$ is a $p \times 1$ vector of soil map unit means and $\boldsymbol{\varepsilon}$ is an independently and identically distributed (iid) random variate with mean zero and variance σ^2. Note that this is a fixed effects model, in which the soil map units are included because the scientist is interested in them; and, having identified, them designs an appropriate scheme on which to sample them. The resulting data are then analysed according to this model. The resulting estimated map unit means, $\widehat{\boldsymbol{\tau}}$, and estimate of the variance of $\boldsymbol{\varepsilon}$, s^2 can then provide a prediction of the value of the soil property at an unsampled site, given the map unit that is delineated there, and an associated prediction error variance (e.g. Leenhardt et al., 1994).

This statistical model is entirely valid, provided that the assumption that $\boldsymbol{\varepsilon}$ is iid is justified by an appropriately randomized sampling scheme (de Gruijter et al., 2006). This is the design-based approach. However, there may be benefits for spatial prediction of soil properties if the spatial dependence of soil variation within map units is modelled statistically, and this is essential where the sample sites have not been selected by an appropriately randomized design (Lark and Cullis, 2004). The model-based approach, which encompasses geostatistical prediction, has been enthusiastically adopted by soil scientists since the seminal work of Burgess and Webster (1980).

In most early soil geostatistics all the soil variation was treated as an autocorrelated random process, but it has been recognized that categorical information, such as conventional soil surveys, and continuous covariates, such as remote sensor measurements, can be combined with geostatistical modelling of the remaining spatial variation. This is the basis of much contemporary work on digital soil mapping (McBratney et al, 2003). Lark et al. (2006) showed how the model in Eq. (1), extended to a linear mixed model, generalized classical geostatistics accordingly. Now the soil property is modelled by

$$\mathbf{Y}(\mathbf{S}) \;=\; \mathbf{X}\boldsymbol{\tau} \;+\; \mathbf{Z}\mathbf{u} + \;\boldsymbol{\varepsilon}, \tag{2}$$

where $\mathbf{Z}$ is a design matrix which associates each observation with a random variable in $\mathbf{u}$. (Note that $\mathbf{Z}$ is $n \times n$ in the usual case where there is no more than one observation at any location in space). The random variable $\mathbf{u}$ is spatially correlated, so it has a correlation matrix $\mathbf{G}$ with the elements on the main diagonal all equal to 1, and off-diagonal elements $\{i, j\}$ taking, in general, non-zero values that depend, under assumptions of second order stationarity, on a parametric function $C(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\psi})$, where the vector $\mathbf{x}_i$ is the location of the ith observation and $\boldsymbol{\psi}$ is a vector of autocorrelation parameters, such as the spatial parameter a of the well-known exponential function

$$C_{\exp}(\mathbf{x}_i - \mathbf{x}_j|a) \;=\; \exp\left\{-\frac{|\mathbf{x}_i - \mathbf{x}_j|}{a}\right\}. \tag{3}$$

It is assumed that the random components have the multivariate normal joint distribution

$$\left[\begin{array}{c} \mathbf{u} \\ \boldsymbol{\varepsilon} \end{array}\right] \sim \mathcal{N}\left(\left[\begin{array}{c} 0 \\ 0 \end{array}\right], \left[\begin{array}{cc} \sigma^2\xi\mathbf{G} & \mathbf{0} \\ \mathbf{0} & \sigma^2\mathbf{I} \end{array}\right]\right), \tag{4}$$

where σ^2 is the variance of $\boldsymbol{\varepsilon}$ and ξ is the ratio of the variance of $\mathbf{u}$ to that of $\boldsymbol{\varepsilon}$. The assumption of an underlying multivariate normal random function is implicit in all standard likelihood estimators, such as the one used in this paper. The data analyst should examine the data to ensure that this is a plausible assumption, perhaps after transformation, but it cannot be absolutely verified. However, it has been observed that likelihood estimators are robust to departures from normality, and that they are optimal estimators by an entropy criterion even in cases where strict normality does not hold (Lark, 2000).

It is worth reflecting on the persistence of the iid component in the linear mixed model, $\boldsymbol{\varepsilon}$ in Eq. (2) and Eq. (4). This 'nugget' variability represents all variation that does not appear spatially correlated over the intervals $\mathbf{x}_i - \mathbf{x}_j$ represented in the data set. This may include measurement error, but the nugget variance can be larger than the known measurement error variance (e.g. Rawlins et al., 2003), indicating that there are substantial sources of variation in the soil operating at larger spatial frequencies (finer scales) than the sampling scheme can resolve.

It is a common assumption in soil geostatistics that the soil exhibits spatially dependent variation at scales bounded at the top of the frequency range (fine scales) by the resolution of the sampling network and not accounted for by the fixed effects in Eq (2). This assumption is generally reasonable, but it is the contention of this paper that it should be formally examined as a matter of course. We can think of the fixed effects in Eq. (2) as representing our soil science knowledge about variable y. This may be the *generalized soil knowledge* of a soil survey if the fixed effects include map units, *specific soil knowledge* if the fixed effect is the prediction from, for example, a process model (e.g. Stacey et al., 2006) or *tacit soil knowledge* that some covariate should be correlated with the variable — Rawlins et al., (2009) used elevation and gamma ray emissions as covariates in a model for soil organic carbon. As we increase the knowledge content of our statistical model so the random effects will become relatively less important. We may also expect that, as our soil knowledge becomes increasingly comprehensive, and as proximal remote sensors become increasingly well-tailored to measuring soil properties, and increase in resolution, so the extent to which we can explain the spatially correlated component of the variation of soil properties should increase. This is illustrated, for example, by Rawlins et al. (2009) who showed the variogram function for the random effects in a model of soil organic carbon decreasing both in sill variance and in the range as more terms were added to the fixed effects part of the model. In short, as we increase the content and sophistication of the soil knowledge of our statistical models so we should examine the possibility that the unexplained variation will not show spatial dependence within the frequency range resolved by the sampling.

A further reason for considering the evidence for spatial correlation in the random component of a mixed model is the subsequent use of the model for spatial prediction. If we assume spatial dependence then the best linear unbiased predictor (BLUP) of the variable at an unsampled site includes a component that is a kriging prediction of the random term. Rather than implement this automatically, we would be best advised to weigh the evidence for spatial dependence, and select an approach to prediction accordingly.

This paper considers the problem of testing the significance of spatial dependence in the random effects of a linear mixed model. The problems are illustrated by simulation, and then the approach is demonstrated in a case study with some data on soil organic carbon.

2. Theory and Simulations

2.1 The linear mixed model and model comparisons in the standard case.

The linear mixed model in Eq (2) is fitted to data, $\mathbf{y}$, by finding an estimate of the random effects parameters, $\boldsymbol{\theta} = [\sigma^2, \xi, \boldsymbol{\psi}]$ that maximizes the residual likelihood:

$$\ell_{\mathrm{R}}\left(\boldsymbol{\theta}|\mathbf{y}\right) \;=\; -\frac{1}{2}\left\{\log|\mathbf{H}| + \log|\mathbf{X}^{\mathrm{T}}\mathbf{H}\mathbf{X}| + (n-p)\sigma^2 + \frac{1}{\sigma^2}\mathbf{y}^{\mathrm{T}}\left(\mathbf{I} - \mathbf{W}\mathbf{C}^{-1}\mathbf{W}^{\mathrm{T}}\right)\mathbf{y}\right\}, \quad (5)$$

where $\mathbf{W} = [\mathbf{X}, \mathbf{Z}]$ and $\mathbf{H} \;=\; \xi\mathbf{Z}\mathbf{G}\mathbf{Z}^{\mathrm{T}} + \mathbf{I}$.

The REML estimate is preferred because it reduces the bias in ordinary maximum likelihood estimation due to error in the fixed effects estimates. The residual likelihood is the likelihood of a variable that is a generalized filtering of the original data such that its maximization provides consistent estimates of the random effects parameters that we require, and it is independent of the value of the unknown fixed effects coefficients (Patterson and Thompson, 1971). One consequence, however, is that the residual likelihood values for alternative models for the same data are comparable only if the models have the same fixed effects structure.

Consider a case where we wish to compare two alternative random effects models, with the fixed effects in common. The first model has the exponential covariance function defined in Eq (3) above. The second is the stable model (Wackernagel, 2003)

$$C_{\mathrm{stable}}(\mathbf{x}_i - \mathbf{x}_j|a) \;=\; \exp\left\{-\frac{|\mathbf{x}_i - \mathbf{x}_j|^{\alpha}}{a^{\alpha}}\right\}, \tag{6}$$

where $0 < \alpha \leq 2$. Our question is whether it is appropriate to use the somewhat more complex stable model. The log-likelihood function for this model, $\ell_{\mathrm{R}}\left(\boldsymbol{\theta}_{\mathrm{stable}}|\mathbf{y}\right)$ is always larger than or equal to that for the simpler model, $\ell_{\mathrm{R}}\left(\boldsymbol{\theta}_{\mathrm{exp}}|\mathbf{y}\right)$ with one fewer parameter so some other criterion is needed. Formal inference can be based on the log-likelihood ratio statistic

$$L \;=\; 2\left(\left\{\ell_{\mathrm{R}}\left(\boldsymbol{\theta}_{\mathrm{stable}}|\mathbf{y}\right) - \ell_{\mathrm{R}}\left(\boldsymbol{\theta}_{\mathrm{exp}}|\mathbf{y}\right)\right\}\right. . \tag{7}$$

Asymptotically under the null model (i.e. when the simpler model holds) this statistic is distributed as χ^2 with degrees of freedom equal to the difference between the number of parameters in the two models, one here, and this test is commonly used in practice to compare linear mixed models (Verbeke and Mohlenbergs, 2000). This distribution applies because the model comparison meets certain conditions (Cox and Hinkley, 1990) which define what is called the standard case for model comparison (Stram and Lee, 1994). Specifically:

i. The models are nested, so that the null model is a special case of the more complex one. This applies for our particular comparison since the exponential model is a special case of the stable model with $\alpha = 1$.

ii. The parameters that take fixed values in the null model are not fixed at boundary values. This applies here because α is fixed at 1 in the null model and can take values in the interval $(0, 2]$ in the stable model.

iii. All remaining parameters in the null model take definite values (i.e. they cannot vary freely without affecting the likelihood). This holds for the exponential covariance function.

To illustrate this I conducted a simulation using the geoR package in R (Ribeiro and Diggle, 2001). I simulated data on a 10×10 square grid using the grf function specifying an exponential variogram with a distance parameter of 5 grid units, a nugget variance, $\sigma^2 = 0.25$ and $\xi = 3$. The only fixed effect was the overall mean, with an expected value of zero. The linear mixed model was then fitted by REML with the likfit function. I specified first an exponential and then a stable covariance function. The maximized likelihoods for the two models were recorded, and the L statistic computed for the model comparison. This was undertaken on a total of 2500 independent realizations. Figure 1 shows the empirical cumulative distribution function of L for this case, with the distribution function for $\chi^2(1)$ superimposed. The two distributions are very close, the 95th percentile for the simulations was 3.1, while the corresponding value of $\chi^2(1)$ is 3.8. The difference is attributable both to sampling error and to the fact that the χ^2 distribution is the asymptotic distribution of L and these values were from a sample of 100.

2.2 Pure nugget vs correlated models: nested but non-standard.

Our objective is to test a model with parameters that describe spatial autocorrelation of the random effects, $\mathbf{u}$, against a pure nugget null model. The question is whether this is a standard case. This is not always a simple question to answer. If there is a parameterization of a null model and an alternative for which all conditions i–iii in section 2.1 hold, then the comparison is a standard case. However, although the conditions do not hold for some particular parameterization of the model they may hold for an alternative parameterization, and so the comparison may be a standard case after all. In short the fact that a particular parameterization that we consider meets the conditions for a standard case is a sufficient but not a necessary condition for the standard case to hold. Haskard (2007) found this in the case of comparisons between an isotropic model (null model) and a geometrically anisotropic equivalent. In the most natural parameterization in which the geometrical anisotropy is expressed as the ratio of the major and minor axes of an ellipse, and the direction of the major axis is a second anisotropy parameter, the case appears to be non-standard since, if the ratio is one then the direction parameter does not take a fixed value. However, Haskard (2007) found by simulation that L for this comparison did have the distribution expected in the standard case, and presented an alternative parameterization of the models in which this is apparent.

For the obvious parameterizations of a nugget null model and a spatially correlated alternative the comparison seems to be non-standard. The argument is as follows.

i. The models are nested, as required. If the autocorrelated model is described by parameters $\boldsymbol{\theta} = [\sigma^2, \xi, \boldsymbol{\psi}]$ then the pure nugget model is a particular case with $\xi = 0$.

ii. However, it is necessary in the autocorrelated model that ξ is non-negative, and so this parameter is at a boundary value in the null model, which violates the second condition for the standard case.

Alternatively one may argue that the pure nugget model is a special case of the general autocorrelated one in which, for example if a spherical autocorrelation function is proposed, the distance parameter is smaller than the shortest interval between any two observations in the data (although larger than the lower bound of zero). However, in this case the parameters ξ and σ^2 do not take definite values, since any combination of values such that $\sigma^2(1 + \xi)$ takes a fixed value will have the same likelihood. This violates the third condition for the standard case. In short, there is no evident way of expressing the pure nugget model as a nested case of a more general variance model for a random effect while meeting all conditions for the standard case, so it seems likely that the distribution of L will follow a mixture of χ^2 distributions.

I tested this conjecture by simulation. I used the grf function in geoR to simulate a pure nugget process on the nodes of a 10×10 square grid, and then fitted linear mixed models with the mean the only fixed effect and with exponential or spherical covariance functions using the likfit function and estimation by REML from random initial values. If the contrasts between each of these spatially dependent alternative models and the nugget null model were standard cases then the distributions of L would be $\chi^2(2)$ since in each case the alternative model has two additional parameters: a (the distance parameter of the exponential covariance function or the range of the spherical covariance function) and ξ. This procedure was repeated to generate 2500 independent realizations of the L statistic for both the spherical versus nugget and exponential versus nugget contrasts. Figure 2 shows the empirical cumulative distribution functions of L for these case, with the distribution function for $\chi^2(2)$ superimposed, from which it seems that, as in other cases (Verbeke and Mohlenbergs, 2006) the distribution function is a mixture of Chi-squareds, including $\chi^2(0)$ for which the distribution funcion is an impulse of height one at zero. Table 1 shows the 95th percentiles for the simulation comparisons, and the equivalent for $\chi^2(2)$. It is clear that the assumption of a $\chi^2(2)$ distribution would result in excessively conservative decisions on average, the 95th percentile for $\chi^2(2)$ exceeds the 99th percentile for both simulated distributions of L. Also reported in Table 1 is the distribution of the log likelihood-ratio statistic when we first select which spatial model (exponential or spherical) to compare according to which has the largest likelihood. The result shows that the effect of this pre-test selection is somewhat anti-conservative, but it is small, and much smaller than the effect of assuming a $\chi^2(2)$ distribution.

On this basis I propose that the sampling distribution of the log likelihood-ratio statistic under the null (nugget) model should be generated by the Monte Carlo procedure used here. To summarize:

i. Generate, at the locations of the sample points, a realization of a random function with a pure nugget variogram.

ii. Fit a linear mixed model with a pure nugget random effect to the data by REML.

iii. Fit alternative mixed models with one or more alternative covariance structures (such as it is intended to use in analysis of real data). These models may be nugget+spherical, nugget+exponential etc.

iv. Identify the best-fitting model from step (iii) and compute the L statistic from the likelihood values for this model and the model fitted at step (ii).

v. Repeat steps (i)–(iv) multiple times to generate a sampling distribution for L under the null hypothesis of a pure nugget process.

It is recommended that the above procedure is used *de novo* since even in the standard case the χ^2 distribution of L is only asymptotic, and so the distribution may be influenced by sample size and configuration. Before demonstrating this approach in a case study, I next report some simulations to evaluate the power of this procedure to reject the null hypothesis of a pure nugget process in a range of circumstances.

2.3 Power to reject the nugget model.

The objective of these simulations was to assess the sensitivity of the log likelihood-ratio as a test of spatial dependence in the random effects of a linear mixed model fitted to data from 100 sample points (with different configurations) within a 10-unit × 10-unit region. The following procedure was used.

i. For the specified sample layout (and in these experiments considering only the nugget+exponential covariance function as an alternative to the null hypothesis of a pure nugget process), use the procedure summarized in the final paragraph of section 2.2 to generate the sample distribution of the L statistic under the null hypothesis, and so a threshold value to reject this null hypothesis with P¡0.05.

ii. Use the grf function in geoR to generate a realization of a random function with a specified covariance function on the same sample points.

iii. Fit two linear mixed models to the simulated data, specifying a pure nugget covariance structure for the first and an nugget+exponential for the second, and compute the L statistic for the comparison. Note whether this statistic exceeds the threshold obtained in step (i), and so whether the null hypothesis may be rejected.

iv. Repeat steps (ii)–(iii) 2500 times, and note the proportion of these simulations for which the null hypothesis is rejected. This is an estimate of the power of the test.

This procedure was followed for a number of cases detailed below.

2.3.1 Square sample grid. In the first set of experiments the sample locations were on a square grid, with grid points on the vertices and edges of the 10-unit × 10-unit region (1.111-unit spacing), see Figure 3a. The distance parameter of the specified model for simulation was either 1 unit or 3 units. The variance of the random variable, $\sigma^2(1+\xi)$ was fixed at 100, and the spatial dependence ratio, $\xi/(1+\xi)$, was set at 0.1, 0.25, 0.5, 0.75, or 0.9. This gives a total of 10 distinct random variables, from each of which 2500 realizations were simulated and modelled. In each case the resulting value of L was compared to the 95th percentile of the empirical sampling distribution for the null model obtained by simulation, and the proportion of occasions on which the null model could be rejected by this criterion was computed.

2.3.2 Augmented square sample grid. This same basic procedure was repeated in a further set of experiments in which the sample sites included 49 locations on a square grid (7 × 7) with points on the vertices and edges of the 10-unit × 10-unit region. An additional 51 points were then added at locations that were optimized, following the procedure of Lark (2002), to minimize the estimation variance of the spatial dependence ratio, assuming that the random variable had an exponential covariance function with a distance parameter of 3 units and a spatial dependence ratio of 0.25. Figure 3b shows the resulting configuration.

2.3.3 Optimized sample array. Finally the same procedure was repeated with 100 sample sites at locations within the 10-unit × 10-unit region that were selected to minimize the same objective function used to optimize the supplementary points in the augemented square sample grid (section 2.3.2 above). The sample array is shown in Figure 3c.

2.3.4 Simulation results The power to reject the null nugget model in favour of the alternative with autocorrelation, estimated by the procedure given in the opening paragraph of section 2.3, is plotted against the spatial dependence ratio for each sampling configuration (Figure 4). Note that in all cases the power to reject the null model declines with the spatial dependence ratio and is about 0.3 or less when the spatial dependence ratio was 0.1. This decline was most marked when sampling on the regular grid and with the shorter distance parameter for the specified covariance function. There were improvements from using the optimized array, particularly the one in which all 100 points were located by the optimization algorithm, and in this case there was little difference in power between cases where the specified distance parameter was 1 and when it was 3. In summary, the power to reject the null model of uncorrelated spatial variation may be small, particularly when the spatial dependence ratio is small and sampling is on a regular array with no additional points to examine variation at short lags. The power of the log likelihood-ratio test depends strongly on the sampling design.

3. A case study with soil data

The soil data reported here were collected at a field at Silsoe in Bedfordshire, England, (52° 0.36′ N, 0° 25.3′ W) as part of a study on within-field spatial variation. The soil was sampled at 100 locations, as shown in Figure 5, and at each location two cores were collected, one from depth 0 – 200mm, and a second from depth 200–800 mm. Soil organic carbon content (SOC) was determined on a subsample from each core following the method of Walkley-Black (Walkley and Black, 1934) and expressed in units of g Carbon per 100g oven-dry soil.

The field had previously been mapped by J.A. Catt from Rothamsted Research, who delineated map units that were allocated to legend classes based on the soil series of the Soil Survey of England and Wales (Clayden and Hollis, 1984). This map has been published elsewhere (Lark et al, 1998). Each soil sample could be unambiguously allocated to a legend unit of this map. The soils in this field are variable. The field lies over a fault between two Cretaceous formations, the Lower Greensand to the north, and the Gault Clay to the South. This solid geology is overlain by colluvium, alluvium and Quaternary deposits. A brief description of the soil series is given in Table 2. As an approximation the soils formed over the Lower Greensand can be designated as Cambisols according to the FAO World Reference base, (IUSS Working Group WRB, 2006) and the soils formed over the Gault Clay as Cambisols and Luvisols. The principal variations among the soils are with respect to texture, caused by differences between their parent materials. Lark et al (1998) showed that the soil map units differed with respect to expected potential soil moisture deficits under different weather scenarios, and with respect to crop yield. Organic inputs to the soil as litter are therefore likely to be variable, and it is known that the dynamics of soil organic carbon are affected by soil texture (e.g. Hassink and Whitmore, 1997) so, a priori, it is plausible that differences between these soil map units will account for significant variation in observed soil organic carbon content. Table 3 provides summary statistics for the data on SOC, including averages within the soil map units, and statistics for the residuals from the map unit averages.

The Monte Carlo procedure summarized in the final paragraph of section 2.2 was used to obtain the sampling distribution of the L statistic under the null (nugget) model with alternatives (i) nugget+exponential covariance, (ii) nugget+spherical covariance and (iii) the best-fitting of two, as judged by the likelihood. Percentiles of these empirical distributions, the threshold values to reject the null hypothesis with P¡0.05 and P¡0.01, are shown in Table 4. Note that these are smaller than the corresponding values for the $\chi^2(2)$ distribution presented in Table 1.

I then fitted linear mixed models to each data set with the grand mean the only fixed effect, and considering pure nugget, exponential and spherical covariance functions in turn. Because only 100 data were available I restricted the models to isotropic ones. The results are shown in Table 5a. Note that for both the upper and lower depth-intervals the exponential model fitted best, as judged by the log likelihood. The log likelihood-ratio for the exponential model compared to the nugget model exceeded the 99th percentile for the simulated distribution for the L statistic computed for the best-fitting model, so the null model can be rejected in both cases with $P < 0.01$.

I then repeated the analyses, but with soil map unit included as a fixed effect. The results are shown in Table 5b. Note that, for both depth intervals, the null hypothesis that the map unit means are all equal can be rejected ($P = 0.038$ for 0–200 mm and $P = 0.009$ for 200–800 mm). At the 0–200 mm depth the random variation exhibits spatial dependence with $L = 11.38$ allowing us to reject the null model with $P < 0.001$. This shows that, at scales between the broad pattern captured by the soil map, and the shortest interval in the sampling scheme (5 m) there are sources of variation in soil organic carbon which cannot be treated as iid random variation. If we require better predictions of SOC than the map unit means then we must sample sufficiently finely to resolve this variation, or identify its origins and develop a proximal sensor of sufficient resolution to account for it. By contrast, at 200–800 mm, while the exponential model fits best by the Likelihood criterion the log likelihood-ratio for the comparison to the null model is small, $L = 1.93$, so we accept the null model ($P > 0.05$). This indicates that no improvement on the map unit mean as a predictor of soil carbon at this depth is possible without sampling at intervals less than 5 m.

4. Discussion

The case study illustrates how the log likelihood ratio can be used to assess the evidence for spatial dependence in the random effects of a linear mixed model for soil properties, by means of empirical sampling distributions of the statistic under the null hypothesis, obtained by simulation. Statistical modelling of soil properties is not a purely numerical exercise but should be undertaken in the light of our understanding of the soil, and to the end of generating new insight. While the parameters of a spatially correlated random effect must always be interpreted with caution it is useful for the soil scientist to know whether the set of covariates and classes included in the fixed effects of the model appears to have accounted for all spatially dependent variation resolved by the sampling scheme, or whether the random variation seems to be caused by processes operating at scales that the sampling can resolve. In this latter case there may be insight into the processes if the random effects are separately mapped by kriging. This was shown by Lark and Webster (2006) in a case study on the height of a geological unit, the Lower Chalk in the Chilterns in south east England. The fixed effects that they considered where polynomial functions of the eastings and northings, which captured the overall geological structure of the unit, the dip. The random variation was then mapped and could be seen to be driven in part by the drainage pattern. In addition to providing insight, an understanding of the spatial structure of the random effects, given evidence that they are spatially correlated, may assist with planning sampling for further study. As in the example in this paper there may be scope to improve predictions with finer-scale sampling, but if there is no evidence for spatial structure of the random effects in the data then this implies that the sampling will have to be on a grid with finer spacing than the shortest interval between existing data points. This may be prohibitive.

Consider the common practice of kriging soil variables from neighbours within a local window, rather than the whole data set. If the random effect is a pure nugget process then kriging is effectively local averaging of a white noise process, which will generate predictions which are a random variable with an imposed correlation structure which is quite spurious. When there is no evidence for spatial correlation of the random effects then the BLUP should be formed with a nugget model. However, the experiments to evaluate the power of the log-likelihood ratio in section 2.3, showed that the evidence for a spatially correlated random effect may be weak for data generated from a spatial process with a large nugget effect. There is often considerable uncertainty about the correlation structure of random effects in linear mixed models, and this is a strong argument for considering Bayesian approaches in which the uncertainty in the variance parameters is accounted for in prediction (Diggle and Ribeiro, 2006). The simulations reported in section 2.3 show that the power of the log likelihood-ratio statistic to reject the null model depends on the sample design as well as on the variance parameters of the random effects.

It is worth considering the relationship between the log likelihood-ratio and Akaike's AIC which is often used to compare alternative variogram models (Webster and McBratney, 1989). The AIC for a particular model, for which the maximized log-likelihood is ℓ and which has κ parameters, is

$$AIC = -2\ell + 2\kappa. \tag{8}$$

The AIC approximates the Kullback-Leibler divergence between the predicted distribution for a new observation, given the data, and its true distribution (Akaike, 1973). If we were comparing two models we would select the one with the smallest value of AIC. This is equivalent, for the case where we compare an alternative model (with κ_a parameters) with a nested null model (with κ_n parameters), to requiring that the log-likelihood ratio

$$L > 2(\kappa_a - \kappa_n).$$

When $(\kappa_a - \kappa_n)$ is smaller than about 6 then the AIC is generally less conservative than L compared to the asymptotic χ^2 distribution for the standard case. However, unlike the likelihood ratio test, AIC is not a formal testing procedure, but rather a rule of thumb for model comparison (Verbeke and Mohlenbergs, 2006).

5. Conclusions

The log likelihood-ratio for comparisons between models with and without spatial correlation of the random effects is not distributed as χ^2 with degrees of freedom equal to the number of additional parameters in the more complex model. The distribution can be approximated by simulation, as detailed in this paper. The power to reject a null model with no spatial correlation may be weak if the sampling scheme is not suitable and the nugget variance is large compared to the spatially correlated variance. There are good practical and scientific reasons to examine the strength of evidence for spatial correlation of the random effects in linear mixed models for soil data, both for prediction and for planning further sampling and measurements of covariates.

Acknowledgements

Soil sampling was undertaken at Silsoe Research Institute by the author and Ms Helen Wheeler as part of work funded by the Ministry of Agriculture, Fisheries and Food (now Defra), and the analyses were completed by staff of the SRI Central Laboratory. Permission was granted by Rothamsted Research to use these data. This paper is published with the permission of the Director of the British Geological Survey.

References

Akaike, H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov B.N, Csaki F. (Eds.), Second International Symposium on Information Theory. Budapest. Akademiai Kiado. pp. 267–81.

Burgess, T.M. & Webster, R. 1980. Optimal interpolation and isarithmic mapping of soil properties. I. The semi-variogram and punctual kriging. Journal of Soil Science 31, 315–331.

Clayden, B., Hollis, J. M. 1984. Criteria for Differentiating Soil Series. Soil Survey of England and Wales Technical Monograph 17, Soil Survey of England and Wales, Harpenden.

Cox, D.R., Hinkley, D.V. 1990. Theoretical Statistics. Chapman & Hall, London.

de Gruijter, J., Brus, D., Biekens, M.F.P., Knotters, M. 2006. Sampling for Natural Resource Monotoring. Springer, Heidelberg.

Diggle, P.J., Ribeiro, P.J. 2007. Model-based geostatistics. Springer, New York.

Haskard, K.A. 2007. An anisotropic Matérn spatial covariance model: REML estimation and properties. Ph.D. Thesis, University of Adelaide. `http://hdl.handle.net/2440/47972`

Hassink, J., Whitmore, A.P. 1997. A model of the physical protection of organic matter in soils. Soil Science Society of America Journal, 61, 131–139.

Hodgson, J.M. 1976. Soil Survey Field Handbook. Soil Survey of England and Wales Technical Monograph 5, Soil Survey of England and Wales, Harpenden.

IUSS Working Group WRB. 2006. World reference base for soil resources 2006. 2nd edition. World Soil Resources Reports No. 103. FAO, Rome.

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

Lark, R.M. 2002. Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. Geoderma 105, 49–80.

Lark, R.M., Webster, R. 2006. Geostatistical mapping of geomorphic variables in the presence of trend. Earth Surface Processes and Landforms 31, 862–874.

Lark, R.M., Cullis, B.R. 2004, Model-based analysis using REML for inference from systematically sampled data on soil. European Journal of Soil Science 55, 799–813.

Lark, R.M., Catt, J.A., Stafford, J.V. 1998. Towards the explanation of within-field variability of yield of winter barley: soil series differences. Journal of Agricultural Science 131, 409–416.

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.

Leenhardt, D., Voltz, M., Bomand, M., Webster, R., 1994. Evaluating soil maps for prediction of soil water properties. European Journal of Soil Science. 45, 293–301.

McBratney, A.B., Mendonça Santos, M.L., Minasny, B. 2003. On digital soil mapping. Geoderma 117, 3–52.

Patterson, H.D., Thompson, R. 1971. Recovery of inter-block information when block sizes are unequal. Biometrika, 58, 545–554.

Rawlins, B.G., Webster, R., Lister, T.R. 2003. The influence of parent material on topsoil geochemistry in Eastern England. Earth Surface Processes and Landforms 28, 1389–1409.

Rawlins, B.G., Marchant, B.P., Smyth, D., Scheib, C., Lark, R.M., Jordan, C. 2009. Airborne radiometric survey data and a DTM as covariates for regional scale mapping of soil organic carbon across Northern Ireland. European Journal of Soil Science 60, 44–54.

Ribeiro, P.J., Diggle, P.J. 2001. geoR: a package for geostatistical analysis. R-NEWS, 1, 15–18.

Stacey, K.F., Lark, R.M., Whitmore, A.P., Milne, A.E. 2006. Using a process model to improve predictions of a soil variable by regression kriging. Geoderma 135, 107–117.

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

Verbeke, G., Molenberghs, G. 2000. Linear Mixed Models for Longitudinal Data. Springer, New York.

Wackernagel, H. 2003. Multivariate geostatistics: an introduction with applications. Springer, Berlin.

Walkley, A., Black, I.A. 1934. An examination of the Degtjareff method for determining organic carbon in soils: Effect of variations in digestion conditions and of inorganic soil constituents. Soil Science 63, 251–263.

Webster, R., Beckett, P.H.T. 1968. Quality and usefulness of soil maps. Nature 219, 680–682.

Webster, R., McBratney, A.B. 1989. On the Akaike information criterion for choosing models for variograms of soil properties. Journal of Soil Science 40, 493–496.

Table 1. Percentiles for L statistic comparing alternative models to a null (pure nugget) for simulated data sets from a pure nugget process on a regular grid. The same percentiles of $\chi^2(2)$ are shown for comparison as this is the asymptotic distribution of L in the standard case.

	Alternative model			
Percentile	Spherical	Exponential	Best-fitting of Spherical or Exponential	$\chi^2(2)$
0.95	2.42	2.84	2.95	5.99
0.99	5.07	5.31	5.39	9.21

Table 2. Principal Soil Series

Series name	Parent Material	Topsoil texture*	Subsoil texture
Lowlands	Colluvium over Lower Greensand	Sandy loam	Sandy clay loam
Hallsworth	Clayey drift over Lower Greensand	Sandy clay loam	Clay loam
Nercwys	Fine loamy drift over Lower Greensand	Sandy loam	Sandy clay loam
Evesham	Clayey drift over Gault Clay	Clay	Clay
Bardsey	Fine loamy drift over Gault Clay	Sandy clay loam	Clay loam
Enborne	Fine loamy alluvium over Gault Clay	Sandy clay loam	Clay loam

*Textural class according to Hodgson (1976).

Table 3. Summary statistics on SOC and residuals from map unit means. SOC is in units of g Carbon per 100g dry soil.

Statistic	SOC 0–200mm	SOC 200–800mm	SOC 0–200mm, residual	SOC 200–800mm
Mean	1.65	1.23	0	0
Median	1.65	1.17	-0.02	-0.09
Minimum	0.43	0.07	-1.10	-1.43
Maximum	3.59	3.67	1.71	1.91
Variance	0.301	0.440	0.229	0.342
Skewness	0.366	0.680	0.460	0.726

Map unit means		
Map unit	SOC 0–200mm	SOC 200–800mm
Lowlands	1.39	0.97
Hallsworth	2.09	1.59
Nercwys	1.89	1.76
Evesham	2.05	1.04
Bardsey	1.88	1.16
Enborne	1.91	1.36

Table 4. Percentiles for L statistic comparing alternative models to a null (pure nugget) for simulated data sets from a pure nugget process on the sample grid from Silsoe.

	Alternative model		
Percentile	Spherical	Exponential	Best-fitting of Spherical or Exponential
0.95	3.57	3.21	4.00
0.99	7.02	6.13	7.23

Table 5. Results for fitting linear mixed models to SOC data for two depths with a). (top) fixed effect the overall mean and b). (bottom) fixed effect the soil map unit mean.

(a)		Depth	
		0–200 mm	200–800 mm
		Log likelihood	
Covariance model for random effects.	Null (nugget)	−83.83	−102.09
	Spherical	−69.77	−100.28
	Exponential	−69.59	−95.12
Model with largest likelihood		Exponential	Exponential
σ^2		0.160	0.290
ξ		1.143	0.511
a		77.3 m	34.3m
L for comparison to null model.		27.58	13.96

(b)		Depth	
		0–200 mm	200–800 mm
		Log likelihood	
Covariance model for random effects.	Null (nugget)	−73.43	−92.31
	Spherical	−68.24	−92.05
	Exponential	−67.74	−91.34
Model with largest likelihood		Exponential	Exponential
σ^2		2.62×10^{-8}	0.22
ξ		38.1×10^{6}	0.664
a		8.4 m	7.3 m
L for comparison to null model.		11.38	1.93

Figure captions.

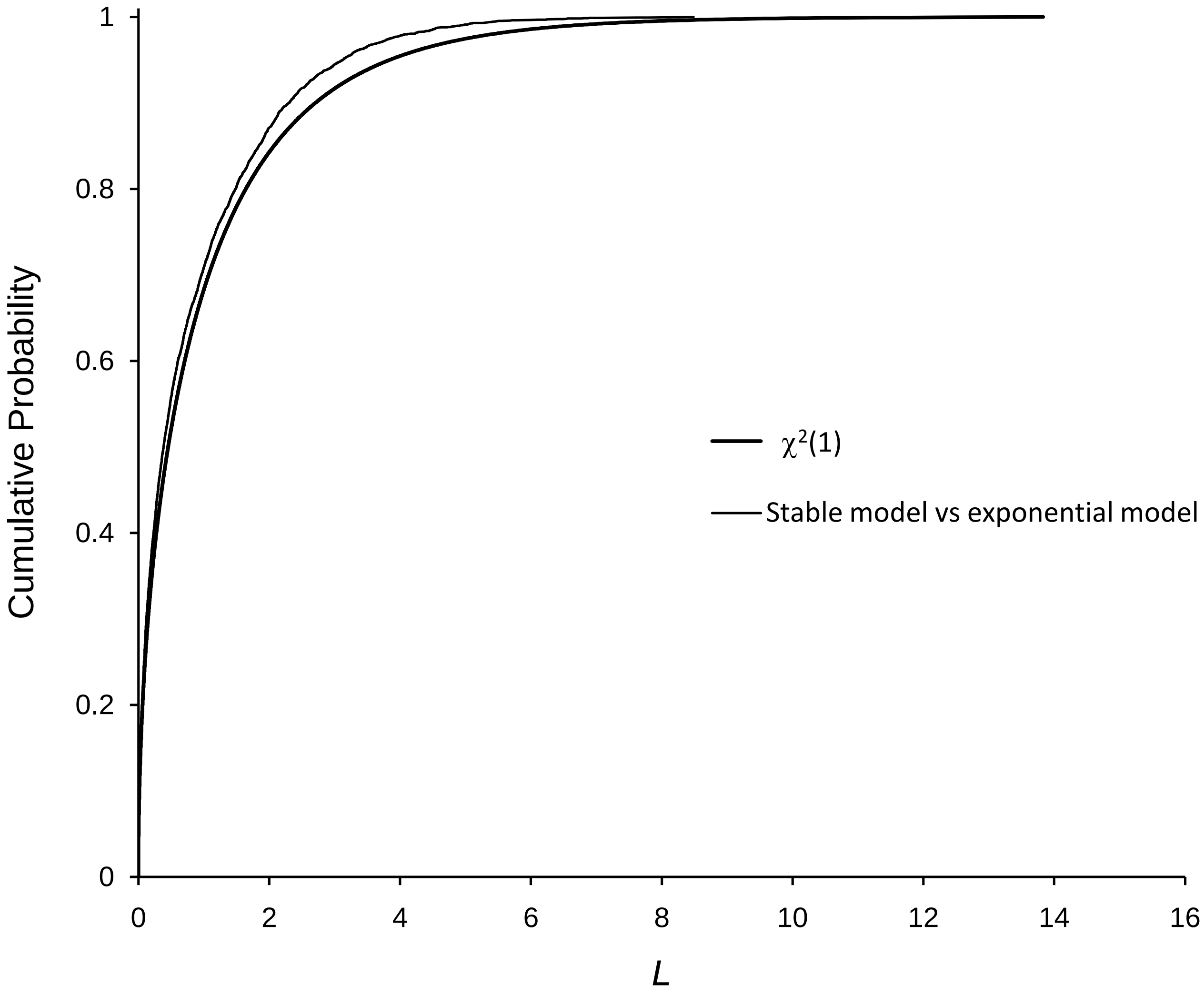
1. Empirical cumulative distribution function of L for a comparison of the stable covariance model with the exponential at 100 locations on a regular grid with data generated from a process with an exponential covariance structure. The distribution function for $\chi^2(1)$ is shown for comparison.

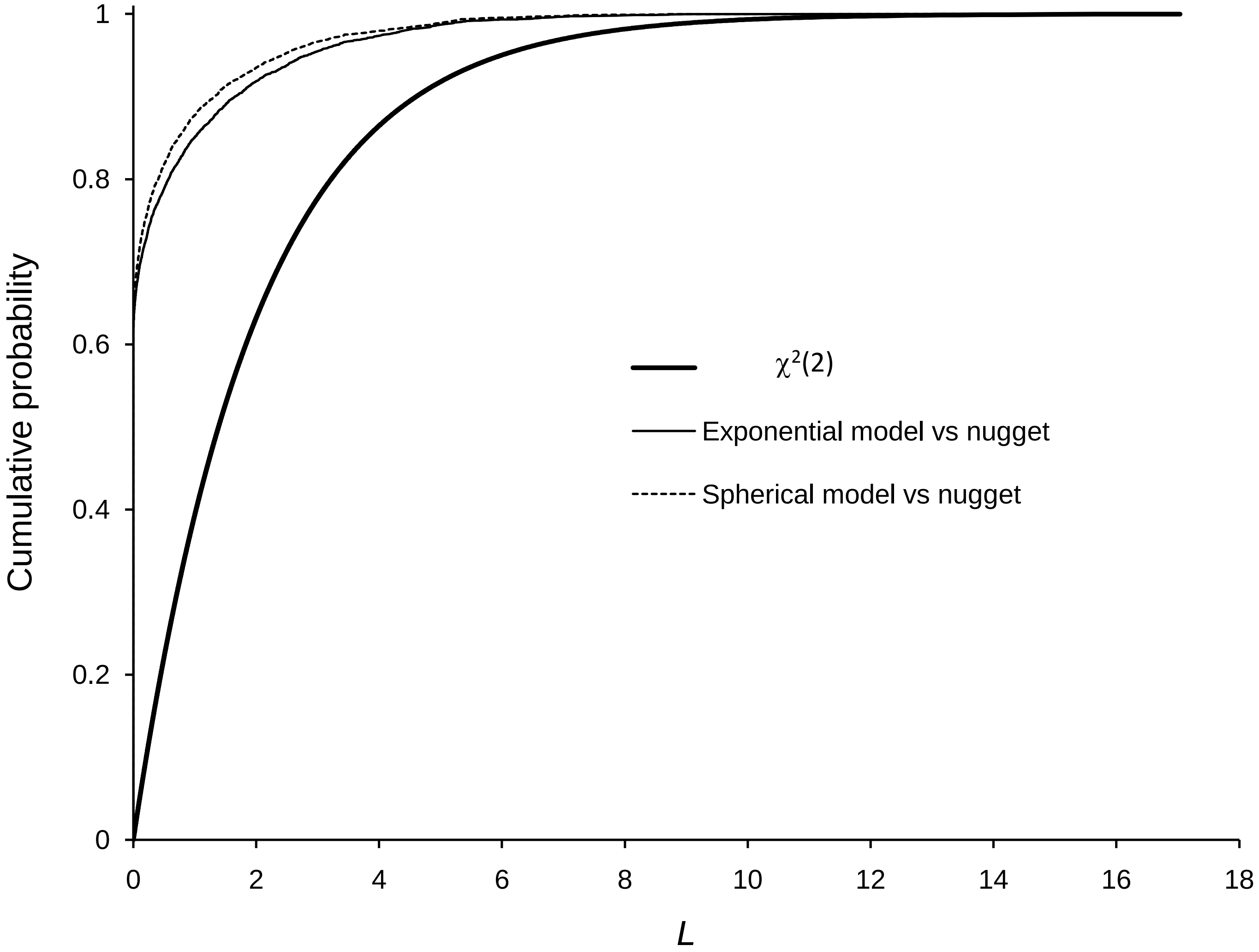
2. Empirical cumulative distribution function of L for a comparison of the exponential covariance model and the spherical covariance model with the pure nugget model at 100 locations on a regular grid with data generated from a process with a pure nugget covariance structure. The distribution function for $\chi^2(2)$ is shown for comparison.

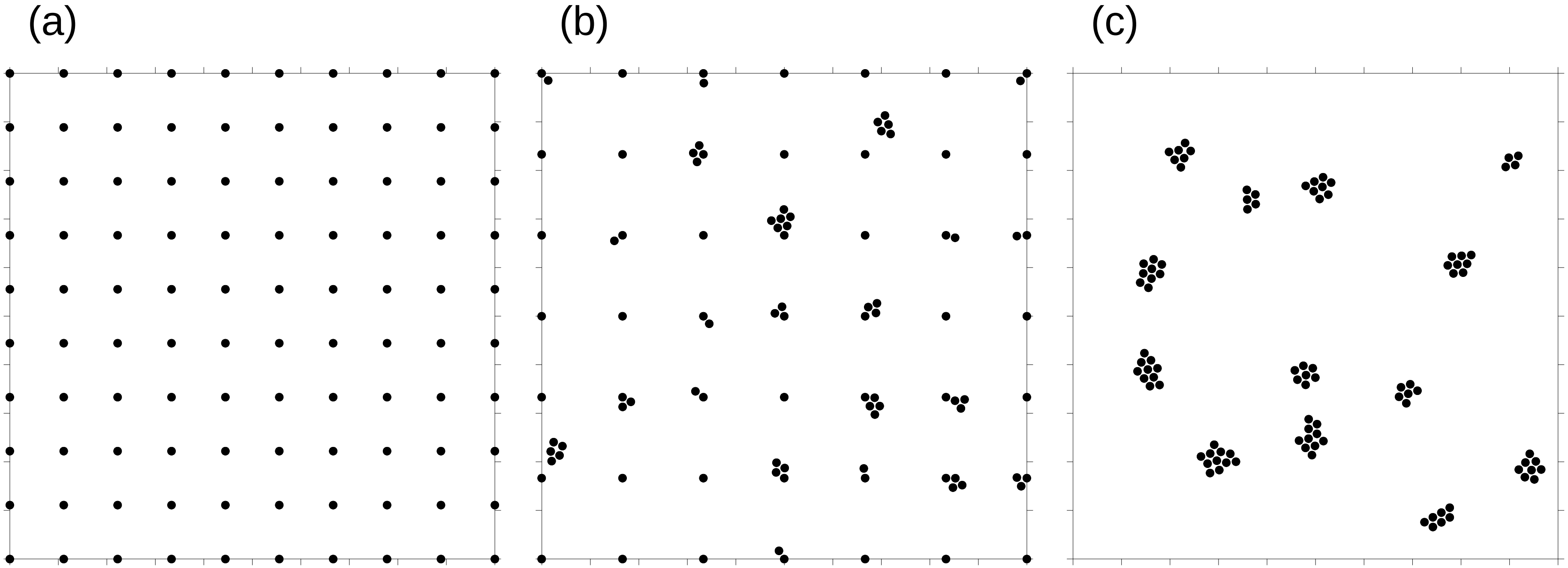
3. Sampling schemes with 100 locations in a 10-unit×10-unit square. a). Regular square grid b). Square grid (49 sites) with 51 sites located to minimize the prediction error variance of the spatial dependence ratio for an exponential model with $a = 3$ and the spatial dependence ratio equal to 0.25. c). 100 locations selected to minimize the same objective function as in (b).

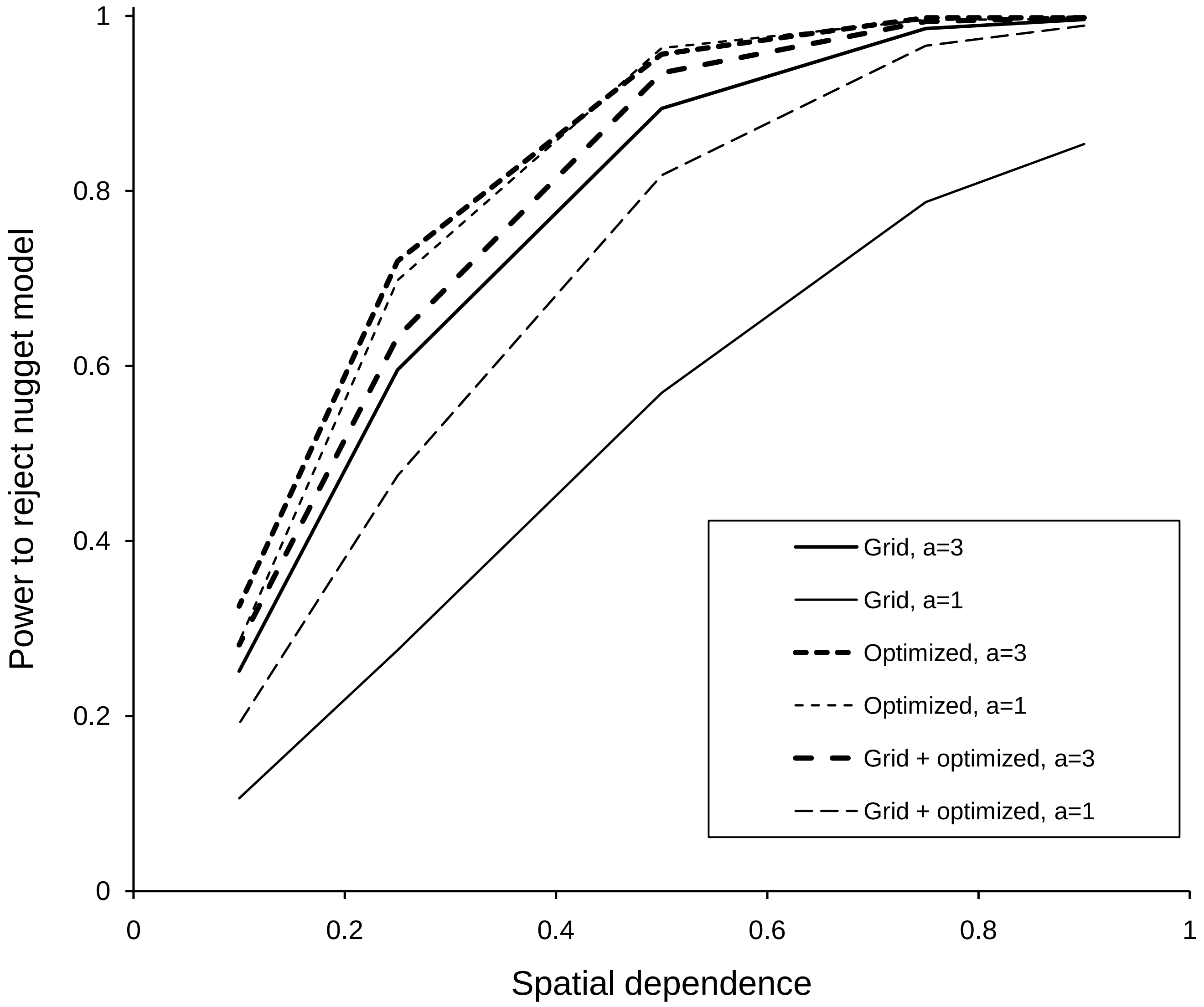
4. Proportion of occasions out of 2500 simulations in which the null (nugget) model is rejected in favour of the exponential plotted against the specified spatial dependence ratio of the simulated process. The three sampling schemes illustrated in Figure 3 are used: grid (Figure 3a), grid with optimized points (Figure 3b) and optimized array (Figure 3c). The specified distance parameter of the simulated process is 1 or 3 units.

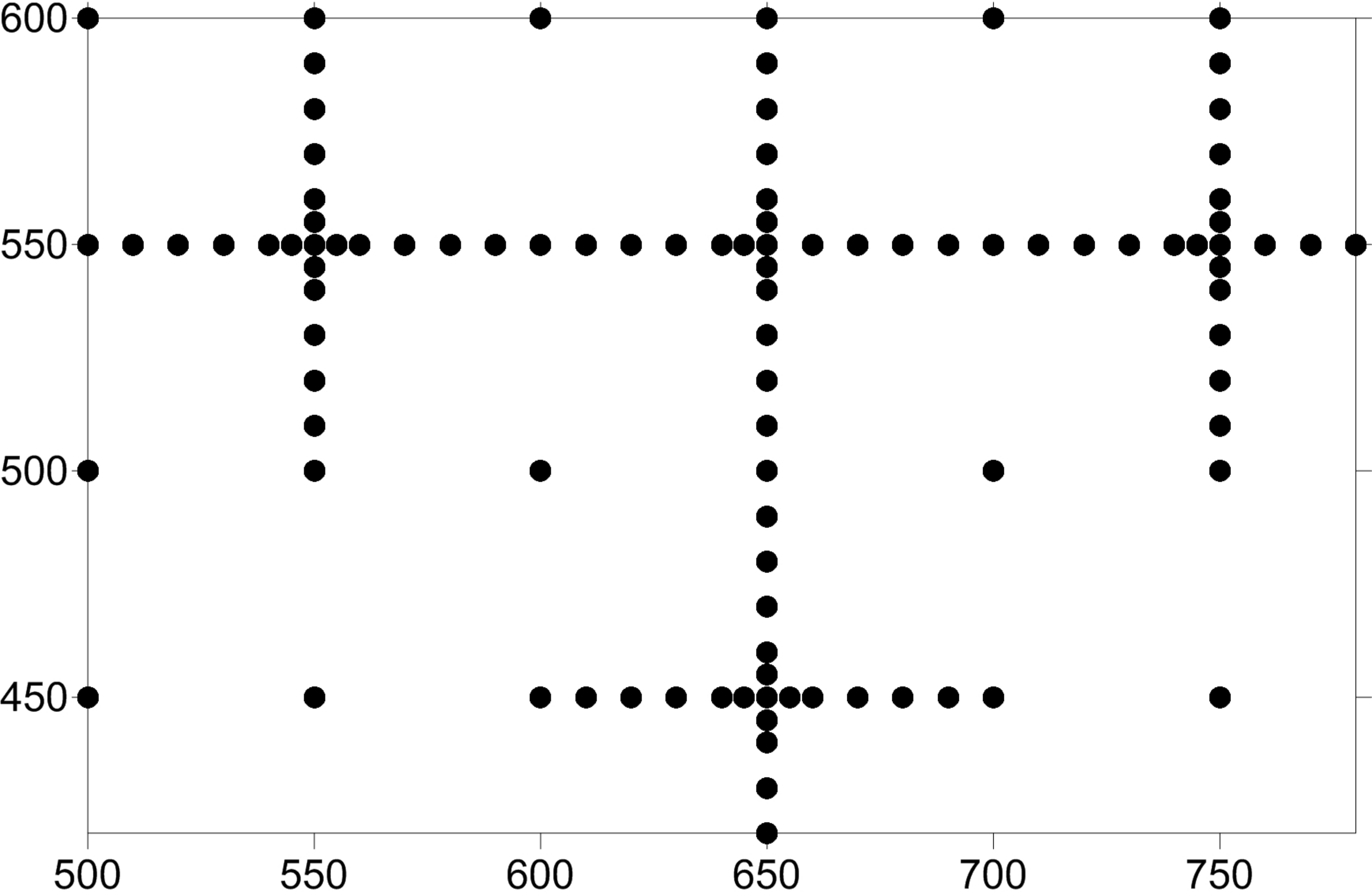
5. Sampling scheme at Silsoe. The coordinates are in metres relative to datum 508000,235000 on the UK Ordnance Survey Grid.


Cumulative Probability
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
16
L
$\chi^2(1)$
Stable model vs exponential model


Cumulative probability
1
0.8
0.6
0.4
0.2
0
0
2
4
6
8
10
12
14
16
18
L
χ²(2)
Exponential model vs nugget
Spherical model vs nugget


(a)
(b)
(c)


Power to reject nugget model
1
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1
Spatial dependence
Grid, a=3
Grid, a=1
Optimized, a=3
Optimized, a=1
Grid + optimized, a=3
Grid + optimized, a=1


600
550
500
450
500
550
600
650
700
750