Some considerations on aggregate sample supports for soil inventory and monitoring

R.M. Lark

British Geological Survey, Keyworth, Nottingham, NG12 5GG

Short title: *Sample support*

Correspondence: R.M. Lark. E-mail: mlark@nerc.ac.uk

Summary

Soil monitoring and inventory require a sampling strategy. One component of this strategy is the support of the basic soil observation: the size and shape of the volume of material that is collected and then analysed to return a single soil datum. Many, but not all, soil sampling schemes use aggregate supports in which material from a set of more than one soil cores, arranged in a given configuration, is aggregated and thoroughly mixed prior to analysis. In this paper it is shown how the spatial statistics of soil information, collected on an aggregate support, can be computed from the covariance function of the soil variable on a core support (treated as point support). This is done via what is called here the discrete regularization of the core-support function. It is shown how discrete regularization can be used to compute the variance of soil sample means, and to quantify the consistency of estimates made by sampling then re-sampling a monitoring network, given uncertainty in the precision with which sample sites are relocated. These methods are illustrated using data on soil organic carbon content from a transect in central England. Two aggregate supports, both based on a 20-m×20-m square, are compared with core support. It is shown that both the precision and the consistency of data collected on an aggregate support are better than data on a core support. This has implications for the design of sampling schemes for soil inventory and monitoring.

Introduction

There is a growing interest in how to sample the soil most efficiently for purposes of inventory and monitoring, spurred by concerns about the impact of human activities on soils and their functions (Arrouays *et al.,* 2009). Among the questions that have been discussed is the choice of sampling design (Papritz & Webster, 1995) and the sources of uncertainty in the resulting estimates (Goidts *et al.,* 2009). Less attention has been paid to the question of what should constitute the support of the basic soil observation.

'Support' is a term from geostatistics. It denotes the size and shape of the volume of material which is analysed to return a single observation in a sample, so the support for a soil observation may be, for example, a vertical cylindrical core of diameter 5 cm and depth 0–15cm. A change of support will result in a change in the statistical properties of soil observations. In practice a support such as a soil core in the example above is so small in comparison to the region of interest that it can be regarded as a point support. The covariance function or variogram of observations on an (effective) point support can be used to compute the statistical properties of observations on a larger support. This process is known as regularization, and is described in standard geostatistical texts (Journel & Huijbregts, 1978; Webster & Oliver, 2009). The question of sample support is discussed briefly by de Gruijter *et al.* (2006). In general increasing the extent of the sample support reduces the contribution of fine-scale variation to our data, this is the regularization effect. It is most readily achieved in soil sampling by bulking.

When we sample soil, and other materials such as water or grain, it may be possible to mix thoroughly a number of specimens (aliquots) from within a specified region, such as an experimental plot, so that the properties of the aggregated material correspond to the average value of the original individual aliquots. This is known as aggregate, bulk or composite sampling. Composite sampling is appropriate for compositional properties of the soil such as its clay or water content or concentrations of elements such as carbon determined by a total element analysis. Exchangeable species can also be determined

from a bulk sample (it is common practice for nutrients) if it can be assumed that the adsorption isotherm is effectively linear over the range of concentrations in the aliquots. Bulk sampling is not generally suitable for soil pH in conditions where significant fragments of carbonate are present in some of the aliquots (Webster & Oliver, 1990), and obviously is not applicable to soil properties that require the structural integrity of soil below some representative elementary volume for laboratory determination (for example, for hydraulic conductivity or parameters of the soil water characteristic curve). De Gruijter *et al.* (2006) discuss sample support and composite sampling separately, but in the case of soil sampling it seems appropriate to define the sample support both in terms of the size and shape of the aliquots, and their spatial distribution. I refer to this as the 'aggregate sample support'. In the case of the National Soil Inventory (England and Wales)(NSI), for example, the aggregate support of the analytical data is 25 cores, each 2.5 cm in diameter and extracted from depth 0–15 cm, collected from a nodes of a square grid of interval 5 m in a 20-m square centred at the nominal sample location (SNIFFER, 2007).

The aggregate sample support varies between different soil sampling schemes. In the United Kingdom we have already seen that the NSI (England and Wales) uses one particular aggregate support. The Geochemical Baseline Survey (G-BASE) of the British Geological survey uses a similar aggregate support for soil: 5 cores (depth 0–15 cm) are collected at the centre and vertices of a 20-m square centred at the nominal sample location and then aggregated (SNIFFER, 2007). The Representative Soil Sampling Scheme (England and Wales) aggregates 20–25 cores collected in a 'W'-pattern across a sample field of no larger than 10 ha. By contrast the Countryside Survey of Great Britain does not undertake aggregate sampling and the sample support for analytical data is a single core (Emmett *et al.,* 2008). Similarly, any analytical datum from the National Soil Inventory of Scotland corresponds to a horizon in a single soil pit (SNIFFER, 2007). The implications of the differences in sample support between these schemes, and the question

of what support is most appropriate, has received little attention.

One reason for this is that, as De Gruijter *et al.* (2006) point out, there is no general theory of composite sampling. Webster & Burgess (1984) considered the use of a single composite sample across a small region to estimate the mean value of soil properties across that region, and gave expressions for the error variance. In this case the aggregate support of a single composite specimen consists of cores drawn from across the region of interest, which might be a field or experimental plot. This does not describe the situation we are concerned with here, in which the region represented by the aggregate support of a single sample is small compared to the overall domain of interest, which may be very large in regional, national or even supra-national soil inventory and monitoring.

Aggregate sample support influences the variability of our basic soil data when we conduct inventory and monitoring across a region, and therefore determines the precision with which we can estimate regional means. It is also likely that sample support will affect the contribution of spatial variation to the sampling error for estimates of temporal change in the soil when monitoring by revisiting a sample network. The aim of this paper is therefore to develop some theory for comparing different aggregate sample supports (including supports in which a single aliquot is collected). Sample supports are compared with respect to the variability of the basic observations made on the support, and so the precision of estimates that we draw from them. They are also compared with respect to the repeatability, site-by-site, of estimates made by re-sampling the soil with error in relocation of the sites, and so the confidence with which we can detect change. Having shown how this can be done, the methods will be applied in order to compare some sampling supports for the measurement of soil organic carbon content, using data collected across a region of lowland England in mixed land-use.

Theory

In this section I first show how one can derive the spatial covariance function of a variable measured on an aggregate support from the covariance function on core-support. This is

a necessary preamble to a demonstration of the effect of sample support on the precision of sampling estimates, and on their site-by-site repeatability.

When we fit covariance functions (or, comparably, variograms) to data on soil and then use these to predict by kriging we are undertaking model-based statistical analysis, in which the random variation of our target variable is assumed to come from an underlying stochastic process, and our data are treated as a realization of a random function which is modelled. This is in contrast to design-based analysis in which we have sampled the soil according to a probability sample design (such as stratified random sampling) and it is this randomized sampling scheme that allows us to analyse our observations as random variables (de Gruijter *et al.,* 2006). However, having fitted a model for a random function we can compute its variance over some region, and can then treat this as the expected value of the variance of the population of values in that region when it is sampled according to a randomized design (Cochran, 1977). This approach was taken by Papritz & Webster (1995) to compare the variances of model and design-based estimates in soil monitoring, and I follow it here. The covariance function on an aggregate support can be used to compute both the variance of observations on that support across a region, and, from this variance, the standard errors of the estimates made from design-based samples of such observations.

I then consider the problem of how repeatable our observations of soil are, site-by-site, in the presence of relocation error. I quantify this by presenting a calculation of the correlation between an observation of the soil, and a repeat observation with relocation error, assuming no underlying change in the soil.

Note that in this paper I assume that all samples are drawn from a two-dimensional space and aggregate sample supports are defined over two-dimensional regions, although the individual aliquots are defined in three dimensions (as with cylindrical cores). The principles, however, would extend simply to aggregation of cores on a transect in one dimension or over volumes in three dimensions.

The covariance function.

In the following sections the observations of a soil variable on a point support are modelled as realizations of a random function, $Z(\mathbf{x})$. We assume that this random function consists of a mean (fixed effect) and a random effect. The mean may be the overall mean of Z across the region of interest, in which case the random effect represents the variation of Z about that mean. Alternatively, we may have divided the region of interest into classes such as soil map units or land-use classes. In this case the mean for $Z(\mathbf{x})$ could be the mean value of Z for the class that occurs at location $\mathbf{x}$, and the random effect is the within-class variation. For simplicity in this section the overall mean is the fixed effect. The random effect is assumed to be a second-order stationary random function which means that it has finite variance and so the spatial covariance function exists:

$$C(\mathbf{h}) \;=\; \mathrm{E}\left[\{Z(\mathbf{x}) - \mathrm{E}\left[Z(\mathbf{x})\right]\}\{Z(\mathbf{x}+\mathbf{h}) - \mathrm{E}\left[Z(\mathbf{x}+\mathbf{h})\right]\}\right], \tag{1}$$

where $\mathbf{h}$ denotes a separation (lag) in space. The covariance declines as the lag distance, $|\mathbf{h}|$, increases and equals zero at lag distances larger than or equal to the range of the covariance function. The *a priori* variance of the random effect is equal to the covariance at lag zero. This is the variance of the variable in a region which is large in comparison with the range of the covariance function. In practice we must fit some appropriate function to describe the covariance of data, and the range (or a related distance parameter) and the *a priori* variance are parameters of this function. One complication that often arises in practice is the nugget effect. There is always some minimum separation, larger than zero, between observations in a real data set and variation that is not spatially dependent at lags larger than this minimum distance cannot be distinguished from spatially correlated variation. As a result the covariance function will appear to converge to some value less than the *a priori* variance as the lag distance decreases, the spatially correlated variance, c_1. The difference between the *a priori* variance and c_1 is the nugget variance, c_0 which is the variance of all components of the random function with spatial dependence over distances smaller than the minimum lag between our observations. A general form of a

model for the covariance function, fitted to data, is therefore

$$\begin{aligned} C(\mathbf{h}) &= c_0 + c_1 , \quad |\mathbf{h}| = 0, \\ &= c_1 \rho(\mathbf{h}) , \quad |\mathbf{h}| > 0, \end{aligned} \tag{2}$$

where $\rho(\mathbf{h})$ is a spatial correlation function such as the spherical

$$\begin{aligned} \rho_{\text{sp}}(\mathbf{h}|a) &= 1 - \left\{ \frac{3|\mathbf{h}|}{2a} - \frac{1}{2} \left(\frac{|\mathbf{h}|}{a} \right)^3 \right\} \quad \text{for } |\mathbf{h}| < a \\ &= 0 \quad \text{for } |\mathbf{h}| \geq a, \end{aligned} \tag{3}$$

where a is presented after the vertical bar because it is a parameter of the correlation function, the range.

The covariances of bulk samples: discrete regularization.

Let $\mathbf{x}_i$ denote the ith sample location, for which a single composite sample is to be formed on an aggregate support. A total of n_i cores is collected at a local array of sites $X_i = \{\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \ldots, \mathbf{x}_{i,n_i}\}$. I assume that the aggregate support is fixed for all sites so $n_i = n_j = n \ \forall i, j$ and $(\mathbf{x}_{i,m} - \mathbf{x}_i) = (\mathbf{x}_{j,m} - \mathbf{x}_j) = \mathbf{a}_m \ \forall i, j; 1 < m \leq n$. I denote the aggregate support by $\mathcal{A} = \{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n | \boldsymbol{\kappa}\}$ where the vector $\boldsymbol{\kappa}$ characterizes the size and shape of a single aliquot.

Let $\breve{Z}_{\mathcal{A}}(\mathbf{x}_i)$ denote a random function, the value of soil property z determined on the material collected on aggregate support $\mathcal{A}$ at location $\mathbf{x}_i$. Note that I follow the usual convention here of putting random functions in upper case and their realizations in lower case. An actual observation of property z on this aggregate support would be written $\breve{z}_{\mathcal{A}}(\mathbf{x}_i)$. I assume that $\breve{Z}_{\mathcal{A}}(\mathbf{x}_i)$ is equal to the arithmetic mean of Z at the locations in the aggregate support

$$\breve{Z}_{\mathcal{A}}(\mathbf{x}_i) = \frac{1}{n} \sum_{m=1}^{n} Z(\mathbf{x}_{i,m}). \tag{4}$$

This ignores any sub-sampling error in extracting material for analysis from the aggregated material, but this error is present in all analysis of field soil samples, regardless of their basic support, and so is not relevant to a comparison between sample supports.

The implication of Equation (4) is that the expectation (mean) of the variable Z within our domain is independent of the aggregate support. This requires that there is nothing in the process of aggregation that introduces bias. We now require a spatial covariance function for the variable on an aggregate support, $\breve{Z}_{\mathcal{A}}$, that is

$$C_{\mathcal{A}}(\mathbf{h}) \;=\; \mathrm{Cov}\left[\breve{Z}_{\mathcal{A}}(\mathbf{x}_i), \breve{Z}_{\mathcal{A}}(\mathbf{x}_i + \mathbf{h})\right], \tag{5}$$

where $\mathbf{h}$ is a lag vector. On the assumption that the variable on point support is stationary in the variance, it is clear from the covariance of two sample means that this expression is given by

$$\begin{aligned} C_{\mathcal{A}}(\mathbf{h}) &= \frac{1}{n^2} \sum_{\mathbf{x} \in X_i} \sum_{\mathbf{x}' \in X_{i+\mathbf{h}}} \mathrm{Cov}\left[Z(\mathbf{x}), Z(\mathbf{x}')\right], \\ &= \frac{1}{n^2} \sum_{\mathbf{x} \in X_i} \sum_{\mathbf{x}' \in X_{i+\mathbf{h}}} C\left(\mathbf{x} - \mathbf{x}'\right), \end{aligned} \tag{6}$$

where $X_{i+\mathbf{h}} = \{\mathbf{x}_{i,1} + \mathbf{h}, \mathbf{x}_{i,2} + \mathbf{h}, \ldots, \mathbf{x}_{i,n} + \mathbf{h}\}$ and $C\left(\mathbf{h}\right)$ denotes the covariance function of the point-support variable Z. In practice we will use a suitable function of the form of Equation (2), fitted to available data on a small enough support (e.g. cores) to be treated as point support.

Equation (6) is directly analogous to the expression for the regularization of the covariance function to a continuous support (Jupp *et al.* 1988). Let $\mathcal{B}$ denote some such support (it might be a square raster pixel in a GIS, for example, that takes the mean value of some variable, such as vegetation cover, over its extent). If $C(\mathbf{h})$ denotes the point-support covariance function of the variable of interest, then the regularized covariance function on support $\mathcal{B}$ is given by

$$C_{\mathcal{B}}(\mathbf{h}) \;=\; \frac{1}{|\mathcal{B}_{\mathbf{s}}||\mathcal{B}_{\mathbf{s}+\mathbf{h}}|} \int_{\mathbf{x} \in \mathcal{B}_{\mathbf{s}}} \int_{\mathbf{x}' \in \mathcal{B}_{\mathbf{s}+\mathbf{h}}} C\left(\mathbf{x} - \mathbf{x}'\right) \; \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}' \tag{7}$$

where $\mathcal{B}_{\mathbf{s}}$ denotes the sample support centred at location $\mathbf{s}$ and $\mathcal{B}_{\mathbf{s}+\mathbf{h}}$ denotes the support with the same size and shape translated to location $\mathbf{s} + \mathbf{h}$; and $|\mathcal{B}|$ denotes the Lebesgue measure of the support (equivalent to its area in two dimensions) and the integrals are over the dimensions of $\mathcal{B}$. The difference between the regularized covariance function and

the expression in Equation (6) is that the former is the covariance of the mean of some variable over a continuous region while the latter is the covariance of the average value of a specific set of discrete observations of the variable on some sample array. For this reason I call Equation (6) the 'discretely regularized covariance function' of the variable, for the specified aggregate support.

The discretely regularized covariance function must be computed from Equation (6) using an available covariance function on a point support, that is one fitted to available data. There may be bias in the regularized function if the lag distances between the individual locations that comprise the support, $|\mathbf{a}_1|, |\mathbf{a}_2|, \ldots, |\mathbf{a}_n|$, are smaller than the shortest distance in the data set from which the point-support covariance function, $C(\mathbf{h})$ is estimated, $|\mathbf{h}_{\min}|$. This is because the fitted model may underestimate or overestimate the covariance at lags shorter than $|\mathbf{h}_{\min}|$. If we wish to evaluate possible aggregate supports then we require covariance functions based on data which include lag intervals shorter than the distances between the aliquots that comprise the aggregate supports of interest. Stein (1999) (page 220) showed that adding a small number of additional points to a regular sample array can substantially improve the modelling of spatial dependence at short distances, and Haskard (2007) showed dramatic improvements in the modelling of short range variation by placing just 10 (out of 100) sample locations at short separations within a sample grid.

Variances of discretely regularized variables.

We have obtained a discretely regularized covariance function for soil data on an aggregate support. Our next objective is to show how we can compute variances of variables measured on this aggregate support. Consider a region $\mathcal{R}$ which we intend to sample on the aggregate support $\mathcal{A}$ at sites selected by simple random sampling. To compute the variance of the resulting sample mean of variable $Z_{\mathcal{A}}$, we require the variance of $Z_{\mathcal{A}}$ in $\mathcal{R}$ according to the covariance model, which we treat as the expected population variance for random sampling. In geostatistics this is called the dispersion variance (Journel &

Huijbregts, 1978) and it can be calculated from the covariance function as:

$$\sigma^2_{\mathcal{A},\mathcal{R}} = C_{\mathcal{A}}(0) - \frac{1}{|\mathcal{R}|^2}\int_{\mathbf{x}\in\mathcal{R}}\int_{\mathbf{x}'\in\mathcal{R}} C_{\mathcal{A}}\left(\mathbf{x}-\mathbf{x}'\right)\ \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'. \tag{8}$$

If the linear extent of $\mathcal{R}$ is large in comparison with the range of the covariance function then the double integral in Equation (8) is negligible and the dispersion variance and the *a priori* variance can be assumed to be equal (Journel & Huijbregts, 1978). Otherwise the second term on the right-hand side of Equation (8) can be calculated most conveniently by a Monte Carlo double integration in which random pairs of locations are drawn from within $\mathcal{R}$ and the average value of the covariance function for the lag interval between them is computed.

It may be that region $\mathcal{R}$ is to be sampled by stratified random sampling. In this case the within-stratum variance is required to compute the standard errors of our estimates. We may distinguish two situations here. In the first, geometrical stratification, the strata are formed by dividing $\mathcal{R}$ into equal subregions, within each of which samples are drawn independently and at random. If one stratum can be represented by region $\mathcal{S}$ then the expectation of the within stratum variance can be computed by substituting $\mathcal{S}$ for $\mathcal{R}$ in Equation (8). Provided that the dimensions of $\mathcal{S}$ are not large relative to the range of $C_{\mathcal{A}}$ the within-stratum variance will be smaller than the dispersion variance for $\mathcal{R}$, wherein lies the benefit of stratification. In the second situation our strata may be categories such as land-use, or soil map units. To obtain the within-stratum variance in this case we require the point-support covariance function for the within-stratum variation. The discretely regularized covariance function of the within-class variation can then be computed, and the expected within-stratum variance is then obtained using Equation (8) evaluating the double integral over the region $\mathcal{R}$.

Some hypothetical examples are presented in Figure 1. Here I considered variables which, on a point support, have a spherical covariance function with range 100 or 500 m and an *a priori* variance of 1.0, of which some varying proportion from 0 to 0.75 corresponds to the nugget variance. I then computed the dispersion variance for these

variables within a 1×1-km block. The calculation was then repeated for the discretely regularized variable, with the support being five aliquots collected at the centre and vertices of a 20-m square then bulked. Note that the dispersion variance on the point support is very close to the *a priori* variance when the range of the covariance function is 100 m, since this is small relative to the dimensions of the block. The dispersion variance is smaller when the range is larger, but the discrepancy decreases as the proportion of the nugget variance increases. In all cases the dispersion variance on the aggregate support is smaller than on the core support. The extent of this reduction in variance by aggregation depends in part on the range of the covariance function, but most dramatically on the relative importance of the nugget variance since this very short-range variation is most susceptible to the regularizing effect of aggregation. If we consider the dispersion variance as the expected population variance for a random sample of the region, it is clear that substantial reductions in the variances of sample means can be achieved by use of an aggregate support. For example, with a nugget variance of 0.25 and a range of 100 m, the variance of the sample mean is reduced by 37% by use of the aggregate support rather than the core support. To achieve this reduction in variance while retaining the core support would require an increase in the number of sample sites of 270%.

Re-sampling and location error.

When monitoring the soil we estimate the change that has occurred in the value of some property over the time period between successive samplings. When our interest is in the change in the spatial mean, then the most efficient sampling design entails revisiting the original sample sites (de Gruijter *et al.,* 2006; Lark, 2009). At the limit the exact sample site cannot be revisited since soil properties are almost always determined destructively by the removal of material for analysis. In practice there is error in the relocation of the sample site, the magnitude of which depends on whether the site is permanently marked or whether it must be relocated by survey from local landmarks or with a GPS. Defra (2003) report studies on the error in locating sample sites for soil monitoring.

A surveyor has visited a site at time t_1 and recorded its location. Let the true location be $\mathbf{x}_1$. At time t_2 the site is relocated as carefully as possible. Let the true location of the identified position be $\mathbf{x}_2$, so the location error is $\mathbf{d} = \mathbf{x}_1 - \mathbf{x}_2$. In practice the surveyor may collect soil at time t_2 from $\mathbf{x}_2 + \boldsymbol{\delta}$ where $\boldsymbol{\delta}$ is a deliberate offset to avoid sampling disturbed ground. I assume that the location error is isotropic (the surveyor is no more likely to err in one direction than another) and that relocation is unbiased, so on average $|\mathbf{d}| = 0$. I assume that the additive effects of sources of location error result in a normal distribution, so that the relocation error is a bivariate normal random variate $\mathbf{D}$ with probability density function $f(\mathbf{D})$ and distribution

$$\mathbf{D} \sim \mathcal{N}\left(0, \sigma_l^2 \mathbf{I}\right), \tag{9}$$

where the mean of zero indicates the lack of bias, and the form of the covariance matrix, with $\mathbf{I}$ the identity matrix, shows that the errors are isotropic, they are uncorrelated and their standard deviation in any dimension is equal to σ_l.

We may characterize the repeatability of a soil monitoring scheme given location error and sampling on a particular aggregate support, $\mathcal{A}$, by calculating the expected correlation between determinations of a soil property on sampling on the aggregate support, and then independently re-sampling on the same support, with location error in each case. We assume that no change occurs between the two samplings, so the differences between the determinations simply reflect spatial variability on the aggregrate support. The mean covariance between the determinations is $C^{1,2}$

$$C^{1,2} = \int_{-\infty}^{\infty} f(\mathbf{D}) C_{\mathcal{A}}(\mathbf{D} + \boldsymbol{\delta}) \mathrm{d}\mathbf{D}, \tag{10}$$

where the integral is over both dimensions of the variate $\mathbf{D}$. This can be scaled to a correlation, $\rho^{1,2}$ by

$$\rho^{1,2} = \frac{C^{1,2}}{C_{\mathcal{A}}(0)}. \tag{11}$$

The stronger this correlation the greater the repeatability of our observations of the soil, site-by-site, on the specified aggregate support.

If there is a substantial nugget effect in the point-support covariance function model which is largely attributable to fine-scale soil variation, then Equation (11) may underestimate the correlation between successive re-samplings of a site because the correlation of the variable over very short distances is underestimated. To compute an upper bound on the correlation $\rho^{1,2}$ I propose that the empirical covariance function in Equation (2) is replaced by

$$\begin{aligned} C'(\mathbf{h}) &= c_0 + c_1 , \quad |\mathbf{h}| = 0, \\ &= c_0 \rho_{\text{sp}}(\mathbf{h}||\mathbf{h}_{\min}|) + c_1 \rho(\mathbf{h}) , \quad |\mathbf{h}| > 0, \end{aligned} \tag{12}$$

where $\rho(\mathbf{h})$ is the fitted correlation function and $\rho_{\text{sp}}(\mathbf{h}||\mathbf{h}_{\min}|)$ is a spherical correlation function with range equal to $|\mathbf{h}_{\min}|$, the shortest distance between observations in the data set from which the empirical model is obtained. Since the spherical correlation function is zero at distances greater than the range this modified covariance function and the fitted one are identical over lag distances larger than $|\mathbf{h}_{\min}|$, and it is assumed that all the variation attributed to the nugget is spatially correlated at distances up to $|\mathbf{h}_{\min}|$. I used a spherical correlation function here because its correlation goes exactly to zero at the range. Other functions with this property (e.g. the circular model) could be used and the choice of function will have a small effect on the computed upper bound.

If it is possible to estimate the independent measurement error for the soil variable of interest, σ^2_{m}, which is a component of the nugget variance, c_0, then Equation (12) may be replaced by

$$\begin{aligned} C'(\mathbf{h}) &= c_0 + c_1 , \quad |\mathbf{h}| = 0, \\ &= \left(c_0 - \sigma^2_{\text{m}}\right) \rho_{\text{sp}}(\mathbf{h}||\mathbf{h}_{\min}|) + c_1 \rho(\mathbf{h}) , \quad |\mathbf{h}| > 0, \end{aligned} \tag{13}$$

I propose that $\rho^{1,2}$ is estimated initially with the discretely regularized form of the fitted covariance function for the target soil property, but that the modified covariance function, Equation (12) or (13) is also used to indicate how much stronger the correlation between site-by-site repeated observations might be if the fine-scale variation is spatially

dependent up to lag $|\mathbf{h}_{\min}|$.

Case study

It has been shown above how the covariance function of a soil property on a point support (approximated in practice by a soil core) can be used to compute the discretely regularized covariance function for observations on an aggregate support. This, in turn, can be used to compute expected values of the variances of the variable on an aggregate support, and to assess the susceptibility of repeated observations of the soil property at a site to relocation error. In this case study I use these methods to calculate, for different sample supports, the variances of means for topsoil organic carbon, obtained by stratified random sampling with land-use classes as strata. I also compute the correlation of repeat samplings of this variable given possible distributions of location errors.

Data and Analyses

The data used here were collected on core support in an agricultural landscape in Bedfordshire, eastern England. The collection of the data is described in detail elsewhere (Haskard *et al.,* 2010, Milne *et al.,* 2011). In summary, the transect was approximately 7.5 km long. The transect started at 508329, 237450 on the UK Ordnance Survey grid (units in metres) and was on a line of bearing 173.5 degrees from grid north, ending at OS grid reference 509182, 229991. There were 256 sample locations at regular intervals (29.45 m) along the transect. To allow analysis of spatial dependence at short distances an additional ten pairs of points were added, each pair comprising one point at 3 m and and one at 6 m along the transect from one of the regular sites. Any variation spatially correlated at distances less than 3 m would therefore contribute to the nugget variance of fitted covariance functions. The soil was sampled at each of the 276 locations to depth 150 mm with a cylindrical gouge auger of internal diameter 44 mm. Milne *et al.* (2011) describe the soils of the transect in more detail. The northernmost point was over the Lower Greensand and the transect intersected the boundary between this formation and

outcrops of the Gault Clay and the Chalk. The southernmost point was at the top of the Chalk escarpment. The soil on the transect is formed in parent materials derived directly from the country rock, and from varied superficial material including alluvium, drift of varied texture and calcareous colluvium below the scarp of the Chalk. Milne *et al.* (2011) also describe land-use along the transect. For purposes of this paper we describe three land-use classes, and assume that these would be used as strata in stratified random sampling of the soil. The classes are arable land (including some land recently set aside, but still under stubble from a recent crop) with 176 observations, woodland (predominantly broadleaf) with 39 observations and uncultivated land (permanent grass, paddock, some waste ground on field margins and some sports grass) with 60 observations.

One sub-sample of the soil from each location was oven-dried to a constant weight to determine the gravimetric water content. Another sub-sample of the soil from each location was air-dried and sieved to pass 2 mm. A sub-sample of the air-dried material was then analysed to determine the total carbon content by combustion in a LECO analyser (LECO CNS 2000 combustion analyser, LECO, St Joseph, Michigan, USA). The carbonate content was determined by the water-filled calcimeter method of Williams (1949) and the organic carbon content (OC) was calculated by subtracting this value from the total carbon content. Soil organic carbon content was then expressed in units of grammes of organic carbon per 100 g dry soil.

One outlying observation was removed from the data set (19 g OC 100 g^{-1} soil). It was very different from the remaining data (the next-largest value was 8.5 g $100g^{-1}$) and would have an undue influence on estimated covariances. Table 1 provides summary statistics on the remaining 275 observations, and on the residuals from the land-use means. These include the octile skewness coefficient (Brys *et al.,* 2003). It was clear that the residuals were reasonably symmetrically distributed, and can plausibly be treated as a realization of a normal random function. The empirical covariance function of the residuals, estimated by the standard methods of moments estimator described by Box &

Jenkins (1976), is shown by the solid symbols in Figure 2. This shows continuity of the covariance down to the shortest lag distance (3 m), and a substantial nugget effect.

A linear mixed model was then fitted to the data (Stein, 1999) by residual maximum likelihood using the lme procedure from the nlme library (Pinheiro *et al.,* 2010) for the R statistical platform (R Development Core Team, 2010). In this model, the land-use was treated as a fixed effect. The empirical covariance function of the residuals suggested that a covariance model with a spatially correlated component (spherical or exponential) and a nugget effect would be appropriate. Both spherical and exponential models were fitted. These can be compared directly with respect to their residual log likelihoods. The log likelihood for the exponential model (-338.5) was larger than that for the spherical model (-341.7) so the exponential model was selected. The estimated fixed and random effects for this mixed model are presented in Table 2. Since the data were on a transect it had to be assumed that the random effect was isotropic. Figure 2 shows the covariance function for the random effects (the covariance of the residuals from the land-use means) with the covariance parameters given in Table 2 (solid line). The modelled covariance is smaller than the empirical covariance function at longer lag distances and the modelled *a priori* variance is larger than the empirical estimate, which is consistent with theory, indicating the bias entailed by estimating the covariance from ordinary least squares residuals (Lark *et al.,* 2006).

I then computed discretely regularized covariance functions for soil organic carbon on different supports, treating the estimated covariance function given in Table 2 as the point-support function. Functions were computed for the NSI (England and Wales) and the British Geological Survey G-BASE soil sample aggregate supports that are described earlier. In both cases the shortest distance between aliquots in the sample support (5 m in NSI and 14 m in G-BASE) is larger than the shortest distance between observations from which the covariance parameters have been estimated (3 m). The difference between the discretely regularized covariance function for these two aggregate supports was negligible

(the *a priori* variances differed by 0.13%). This is of interest because it suggests that collecting as many as 25 individual cores from a 20-m square may not be justified (unless it is necessary to ensure sufficient soil for the planned analyses). To investigate this further I computed the *a priori* variance for data on an aggregate support based on a 20-m square with varying numbers of cores. The variances are plotted against the number of cores in Figure 3, which also shows the disposition of cores within a single square. This confirms that the variance drops rapidly as the number of aliquots is increased to five, but adding further aliquots has little effect.

I then computed dispersion variances by Monte Carlo integration for soil organic carbon on the point support and G-BASE aggregate support within square domains with sides of various lengths between 1 and 10 km. These are plotted on Figure 4, including the *a priori* variances which are very close to the dispersion variances for regions length 5 km or more. The *a priori* variance of the variable on an aggregate support is 32% of that on the point support. The dispersion variances on the aggregate support within a 1-km square block is 36% of that on the point support.

Assume that a sample is drawn from a region of 10-km×10-km or larger, with soil variability comparable to the landscape investigated here. Stratified random sampling is used with the land-use classes as strata. The *a priori* variances of the point-support and aggregate support covariance functions computed from the estimated parameters in Table 2 would be the expected pooled within-stratum variance, σ^2_{w} (there are not sufficient data here to estimate separate variance parameters for the different classes). The standard error of the mean SOC estimated from N observations distributed in proportion to the areas of the different strata would be $\sqrt{\frac{\sigma^2_{\mathrm{w}}}{N}}$. If we wanted the 95% confidence interval on an estimate of the mean to be approximately ± 10% of the mean (which is 2.7 g $100\mathrm{g}^{-1}$ soil) then calculations show that we would require about 62 samples on a core support, but because of the smaller variance on the aggregate support only 42 aggregate samples would be required. This would be a substantial saving of field effort and analytical costs.

I used Equations (10) and (11) to compute the expected correlation between data obtained by two samplings of the same set of locations, assuming that the location error $\mathbf{D}$ is normally distributed with mean zero and different standard deviations, and that the offset $\boldsymbol{\delta}$ to avoid re-sampling disturbed sites is 10 cm. The point-support covariance function with parameters in Table 2 was used. The results are plotted in Figure 5 for point, G-BASE and NSI support. An upper bound for the correlation was also obtained by substituting a spherical covariance function for the nugget as in Equation (12) with $|\mathbf{h}_{\min}| = 3$ m, and this is also shown in Figure 5.

Defra (2003) reports estimates of relocation error in revisiting soil sampling sites. On enclosed land it was estimated that the relocation error was less than 10 m in 61% of trials. If the relocation error is assumed to be bivariate normal then this implies a standard deviation in any one dimension of about 7 m. On open land it was estimated that the error was less than 10 m in 33% of cases, which implies a standard deviation in any one dimension of about 11 m. When the standard deviations of the location error are of this magnitude the difference between the calculated correlation of the sampled and re-sampled observations and the upper bound of this correlation are negligible. The calculated correlations with the standard deviation (one dimension) of 7 m were 0.62, 0.89 and 0.89 for the core, G-BASE and NSI supports respectively, and were very little different for a standard deviation of 11 m (0.61, 0.87, 0.88).

Discussion

A geostatistical analysis allows us to make some plausible inferences about the relative merits of different sample supports for soil inventory and monitoring, provided that we have robust spatial covariance functions for the soil variable of interest from data sets which allow us to model spatial dependence over distances less than the intervals between aliquots of any proposed aggregate support. There is a general awareness that robust planning of soil inventory and sampling requires some exploratory data on soil variability, and this paper shows that information on fine-scale variation is also needed. This should

be a priority for future work on soil variability for planning soil surveys. Short-range variability of soil properties may differ markedly between soils on different parent materials and with different histories of land-use. It is therefore unrealistic to expect that a general purpose covariance model will describe the effects of aggregation on the statistics of soil data across a country or even a large region. This is true of any decisions on sampling strategy based on observed statistics, some degree of generalization is unavoidable. In practice two options are possible. One could identify areas where the fine-scale variation of the soil is likely to be largest, and sample that region to obtain a covariance function at fine scales to plan the sampling support. This would ensure that the precision of measurements in the most variable regions was adequate. Alternatively, one might obtain covariance functions for general regions which are expected to differ in their variability (for example, lowland arable soils and upland grassland) and plan different sample supports for these regions so that the precision in each is similar.

There are potentially large differences between the *a priori* variances of soil data on point and aggregate supports, since in the latter case short-range variation is removed by the process of bulking. The extent of this advantage depends on the spatial covariance function of the variable on the point support. Nonetheless, it is clear that there are potential advantages in using an aggregate support when the objective is to sample to characterize a large region. While we cannot generalize from the results presented here on soil organic carbon, from one particular data set, it is notable that the *a priori* variance on the aggregate support is some 30% less than that on a core support, and about 30% fewer samples were required to meet a reasonable quality standard for estimating regional mean soil carbon content by stratified random sampling.

Given the interest in soil monitoring, the results on the effect of sample support on the repeatability of sampling are important. These show considerable improvements in the correlation between independent determinations of soil properties over sites when an aggregate support is used. This is plausible since, with even quite large relocation errors,

the region of the aggregate support for the baseline and re-sampled observations will often overlap, and aggregation reduces the short-range variation which contributes most to the uncertainty in site-by-site comparisons over time. Figure 5 shows that the differences in correlation can be large even when the relocation standard deviation is small, which suggests that this is an significant consideration even as the performance of GPS or other technology to aid relocation improves.

The aggregate sample requires more effort to collect at the local site than a single core. The local grid must be marked out, and the samples collected, physically mixed and sub-sampled. However, it is likely that the additional cost of these operations within each sampling site will be less than that of adding additional sites to a randomized scheme, with the administrative overheads, travel and analytical costs that each additional site entails. It is also of interest in this case that the benefits of increasing the number of aliquots within a 20-m square beyond five were negligible. By calculating the *a priori* variance of observations on aggregate supports with different numbers of aliquots we can make a rational decision as to how many are required to achieve target precision (although there must also be enough to provide sufficient material for the planned analyses and for archiving).

As noted earlier, there is considerable variation in the sample support used in different schemes for soil inventory and monitoring, even within the United Kingdom. These results suggest that it is advantageous to use an aggregate support where this is possible for the soil properties of interest. Is it appropriate for existing surveys to change the support that they use? A change in the support of soil data can, in principle, influence all its statistics. It would clearly be undesirable to change the support of soil data if this would change the mean. If the depth in the soil from which individual aliquots are extracted remains unchanged then a change of support should not affect the mean of a compositional property expressed gravimetrically such as organic carbon content or available nutrients. Provided that the size and shape of the aliquots (as determined by sampling

tins or augers) as well as sample depth are unchanged then introducing an aggregate support would have no effect on the mean of volumetric properties such as porosity or bulk density. A change in the variance of soil data caused by a change in support need not cause problems for the statistical analysis of the resulting data and their comparison with earlier observations on a different support. There are quite standard expressions, for example, to compute a standard error on the difference between two independent samples of a variable when the samples cannot be assumed to have the same variance (Snedecor & Cochran, 1989).

Conclusions

To conclude, provided that we have a sound model of the spatial covariance of a soil property on point support, it is possible to compute the discretely regularized covariance function for that same property on a range of aggregate supports. This function can be used to compute the variance of the soil property on those supports within regions of any size or shape, and to calculate how consistently the soil can be re-sampled on the particular support, given relocation error.

In the case of soil organic carbon in a lowland environment, it was shown that the variance of observations on the aggregate supports used by the National Soil Inventory (NSI) of England and Wales, and the British Geological Survey's Geochemical Baseline Survey (soils) is substantially smaller than on a single core support, and that the consistency of re-sampling is also greater. To form robust conclusions across a range of conditions and soil properties would require further sampling to allow us to model the spatial covariances of these properties at fine (within-support) scales.

Acknowledgments

This paper is published with the permission of the Director of the British Geological Survey. I acknowledge the permission of Rothamsted Research to use the data on soil carbon which I collected while in its employment. I am grateful to the associate editor

533 and two referees for helpful comments on this paper.

References

Arrouays, D., Bellamy, P.H & Paustian, K. 2009. Soil inventory and monitoring: current issues and gaps. *European Journal of Soil Science*, **60**, 721–722.

Box, G.E.P. & Jenkins, G.M. 1976. *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco.

Brys, G., Hubert, M. & Struyf, A. 2003. A comparison of some new measures of skewness. In: Developments in Robust Statistics (eds R. Dutter, P. Filzmoser, U. Gather and P.J. Rousseeuw), pp. 98–113. Physica-Verlag, Heidelberg.

Cochran, W.G. 1977. *Sampling Techniques*. 3rd Edition. John Wiley & Sons, New York.

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

Defra. 2003. *Sampling strategies and soil monitoring: Report SP0514*. Defra, London.

Emmett, B. A., Frogbrook, Z. L., Chamberlain, P. M., Griffiths, R., Pickup, R., Poskitt, J. *et al.* 2008. *Soils manual. Countryside Survey Technical Report no 03/07*. Centre for Ecology and Hydrology, Lancaster.

Goidts, E., van Wesemael, B. & Crucifix, M. 2009. Magnitude and uncertainties in soil organic carbon (SOC) stock assessments at various scales. *European Journal of Soil Science*, **60**, 723–739.

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`

Haskard, K.A., Welham, S.J., Lark, R.M. 2010. A linear mixed model with spectral tempering of the variance parameters for nitrous oxide emission rates from soil across an agricultural landscape. *Geoderma*, **159**, 358–370.

Journel, A.G. & Huijbregts, C.J. 1978. *Mining Geostatistics*. Academic Press, London.

Jupp, D.L.B., Strahler, A.H. & Woodcock, C.E. 1988. Autocorrelation and regularization in digital images. I. Basic theory. *IEEE Transactions on Geoscience and Remote Sensing*, **26**, 463–473.

Lark, R.M. 2009. Estimating the regional mean status and change of soil properties: two distinct objectives for soil survey *European Journal of Soil Science* **60**, 748–756.

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.

Milne. A.E., Haskard, K.A., Webster, C.P., Truan, I.A. Goulding, K.W.T. & Lark, R.M. 2011. Wavelet analysis of the correlations between soil properties and potential nitrous oxide emission at farm and landscape scales. *European Journal of Soil Science* **62**, 467–478.

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

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D & the R Development Core Team. 2010. *nlme: Linear and Nonlinear Mixed Effects Models.* R package version 3.1-97.`http://www.R-project.org/`.

R Development Core Team. 2010. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. `http://www.R-project.org/`.

SNIFFER, 2007. *National Soil Monitoring Network: Review and Assessment Study.* SNIFFER, Edinburgh, UK. LQ09, June 2007.

Snedecor, G.W., Cochran, W.G., 1989. *Statistical Methods.* 8th Edition. Iowa State University Press, Ames, Iowa.

Stein, M.L. 1999. *Interpolation of Spatial Data: some Theory for Kriging.* Springer, New York.

Webster, R. & Burgess, T.M. 1984. Sampling and bulking strategies for estimating soil properties in small regions. *Journal of Soil Science*, **35**, 127–140.

Webster, R. & Oliver, M.A. 1990. *Statistical Methods in Soil and Land Resource Survey.* Oxford University Press, Oxford.

Webster, R. & Oliver, M.A. 2009. *Geostatistics for Environmental Scientists.* 2nd Edition John Wiley & Sons, Chichester.

Williams, D.E. 1949. A rapid manometric method for the determination of carbonate in soil. *Soil Science Society of America Proceedings*, **25**, 248–250.

Table 1. Summary statistics for data on soil organic carbon from the Bedfordshire transect (after removal of one outlier).

	Soil organic carbon /g carbon $100g^{-1}$ soil.	Residual from land-use mean /g carbon $100g^{-1}$ soil.
Mean	2.66	0.00
Median	2.39	0.00
Standard deviation	1.30	1.02
Skewness	1.84	0.27
Octile skew	0.17	0.004
Minimum	0.12	-2.89
Maximum	8.52	3.97

Table 2. Estimated parameters for a linear mixed model fitted to data on soil organic carbon from the Bedfordshire transect.

Fixed effects	Mean soil organic carbon content /g $100g^{-1}$ soil
Arable	2.20
Wood	4.11
Uncultivated	3.04

Random effects	Parameter values
Correlation function type.	Exponential
c_0	0.392 g^2 100 g^{-2}.
c_1	0.738 g^2 100 g^{-2}.
a	215.8 m

Figure Captions.

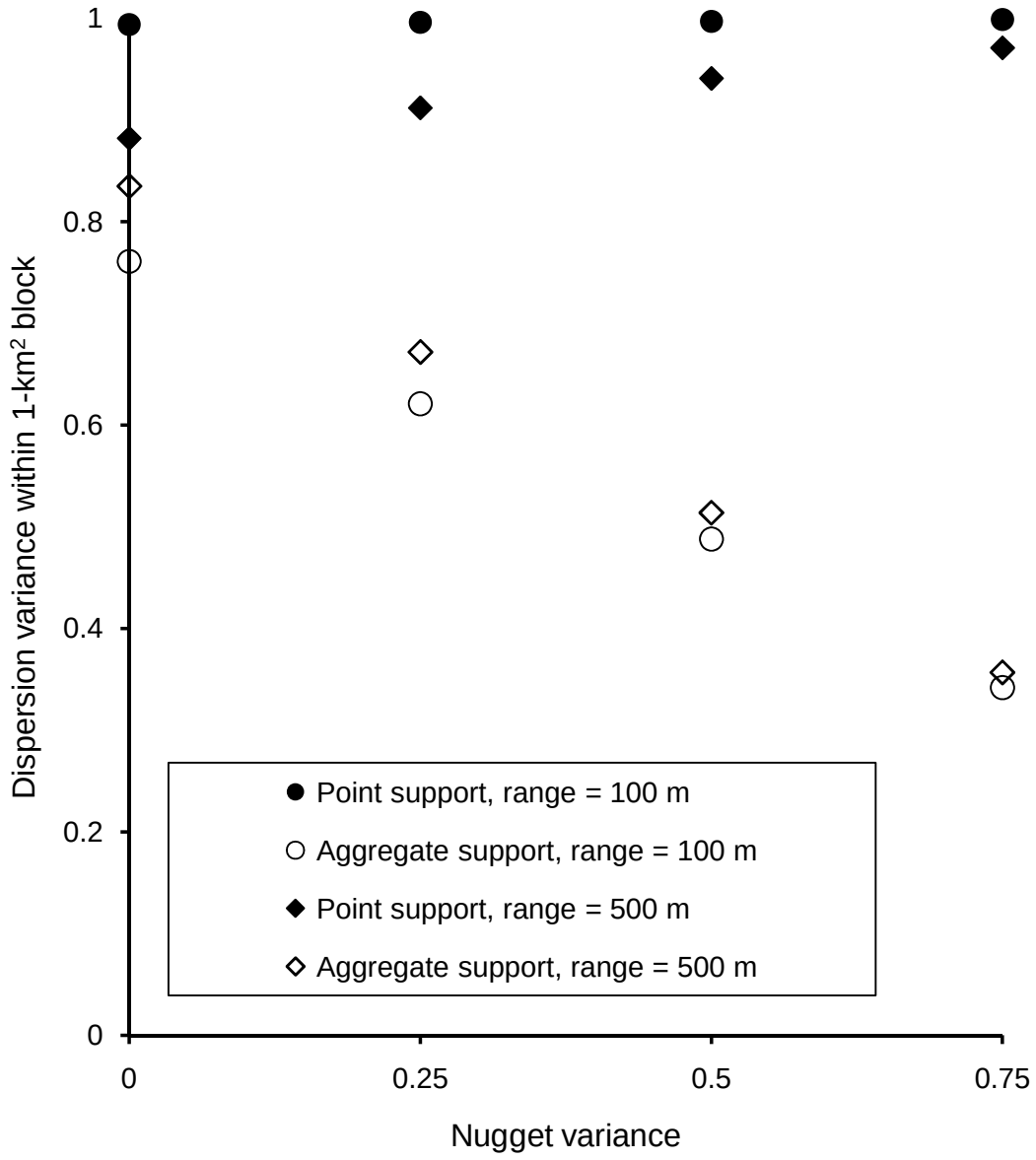
1. Dispersion variances within a 1×1-km block for data with point support (solid symbol) or aggregate support (open symbol) on random functions with a spherical variogram, *a priori* variance 1.0, differing nugget variances (abscissa) and a range of 100 or 500 m.

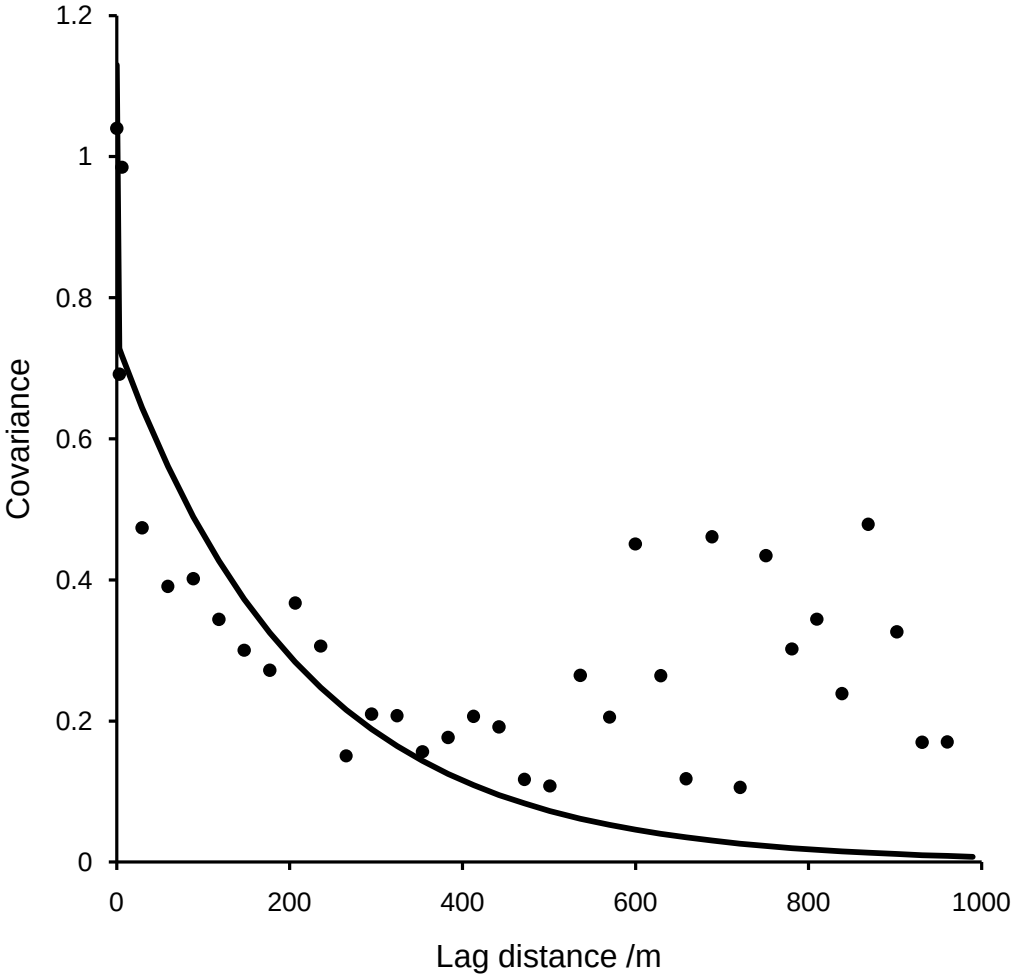
2. Empirical covariance function (solid symbols) of soil organic carbon residuals from land-use mean. The solid line is the (point-support) exponential covariance function with parameters estimated by REML for the linear mixed model for soil organic carbon with land-use as a fixed effect.

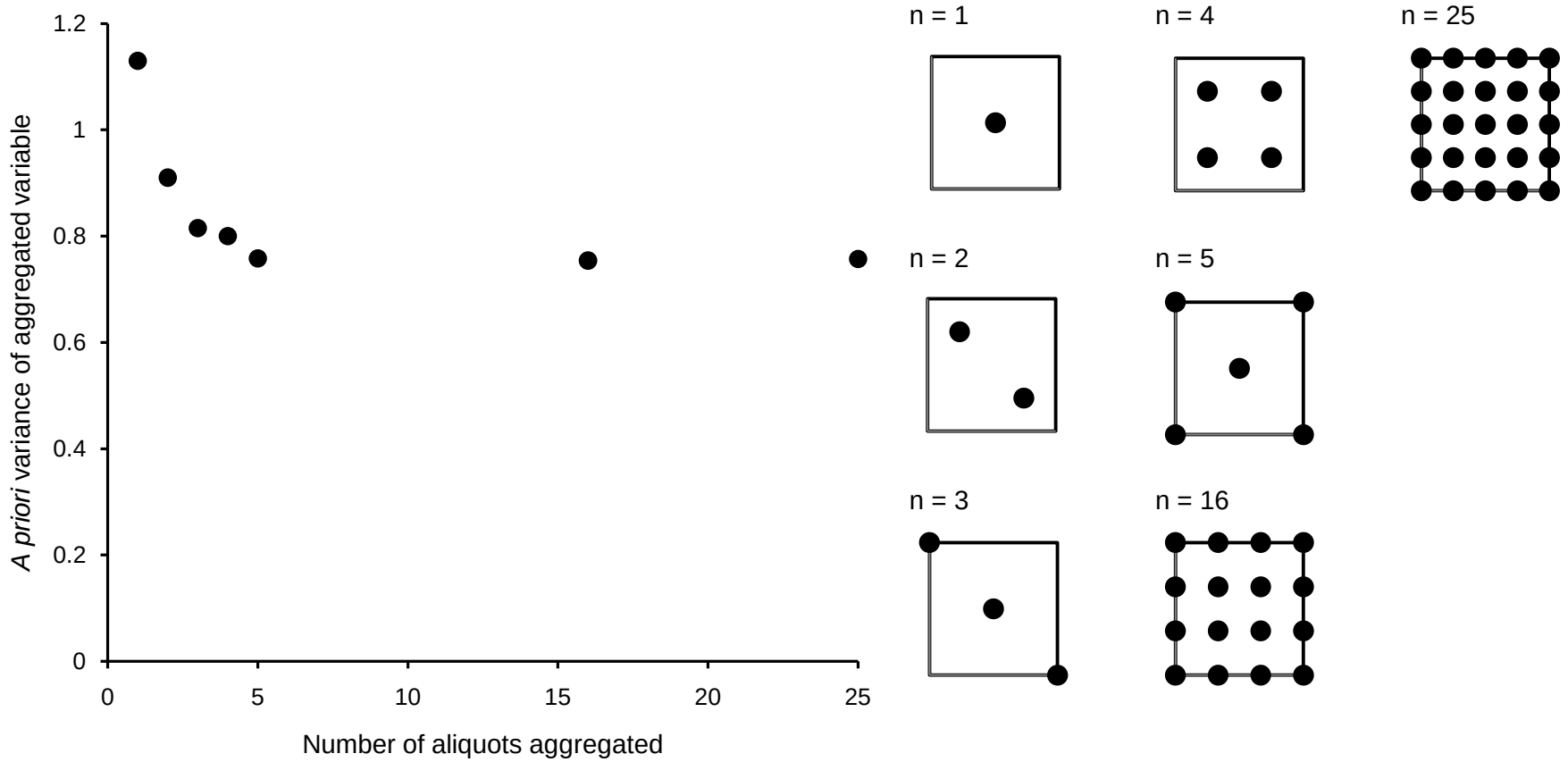
3. Expected *a priori* variances of measurements of soil carbon for measurements on seven different sample supports (illustrated). Each support is based on a 20×20-m square and has differing numbers of aliquots (indicated by solid symbols).

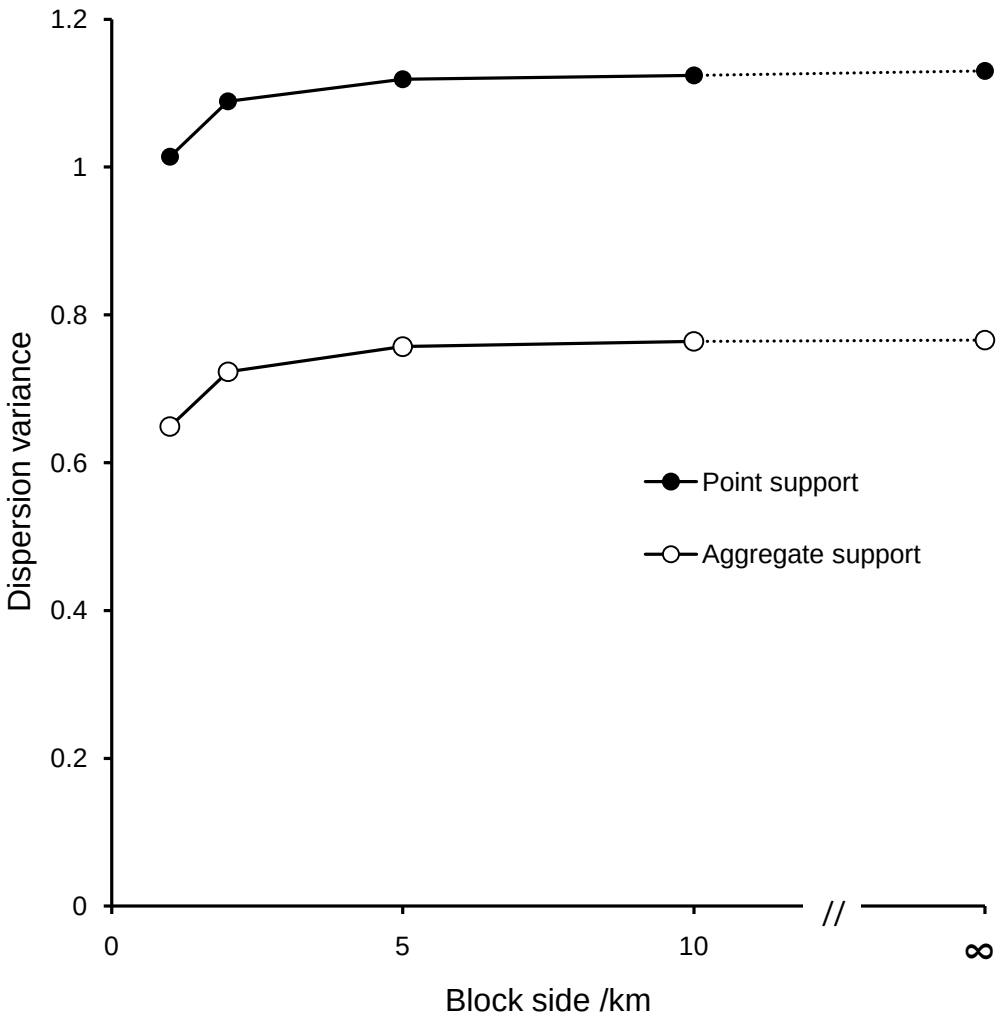
4. Dispersion variances for soil organic carbon (within land-use) on point support or aggregate support (G-BASE) within square blocks with differing lengths.

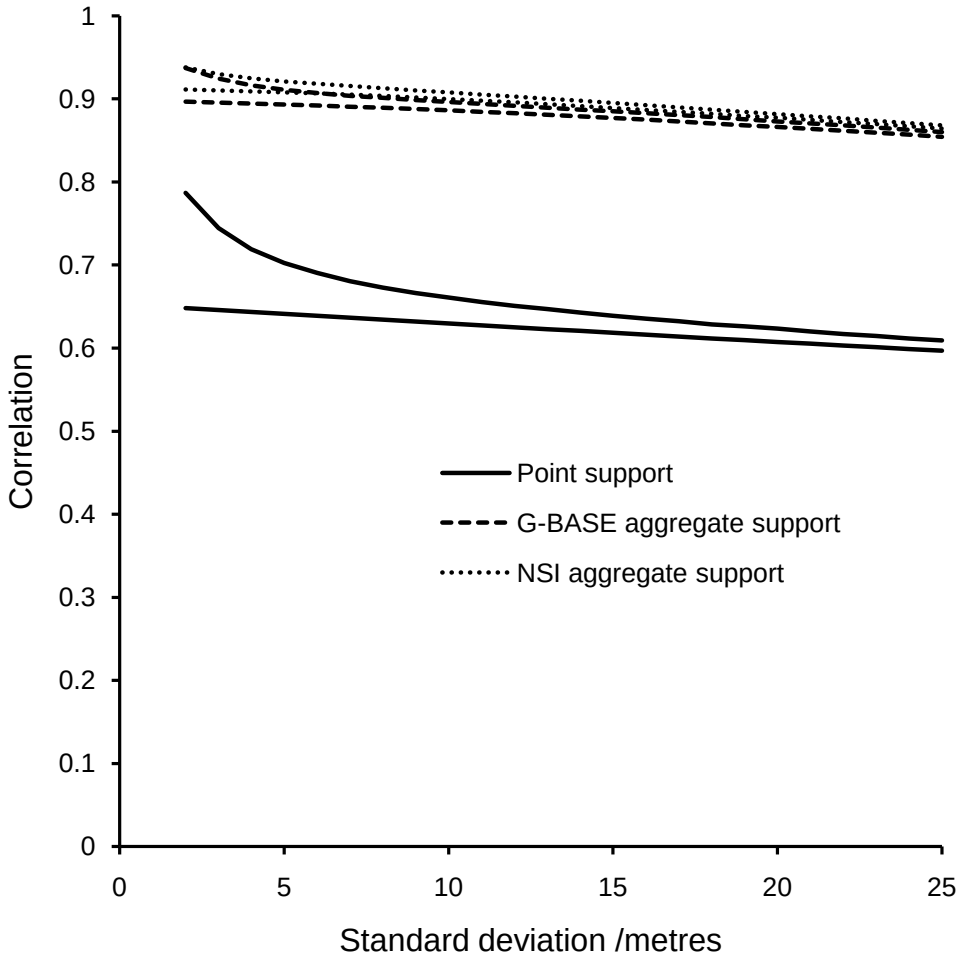
5. Correlation between two independent re-samplings of soil carbon on point and aggregate supports plotted against standard deviation (in any one dimension) of the relocation error. For each support the lower line is the correlation calculated from the fitted covariance function, and the upper line is an upper bound on the correlation calculated with the covariance function given in Equation (12).


1
0.8
0.6
0.4
0.2
0
Dispersion variance within 1-km2 block
0
0.25
0.5
0.75
Nugget variance
Point support, range = 100 m
Aggregate support, range = 100 m
Point support, range = 500 m
Aggregate support, range = 500 m


1.2
1
0.8
0.6
0.4
0.2
0
Covariance
0
200
400
600
800
1000
Lag distance /m


1.2
1
0.8
0.6
0.4
0.2
0
0
5
10
15
20
25
A priori variance of aggregated variable
Number of aliquots aggregated
n = 1
n = 4
n = 25
n = 2
n = 5
n = 3
n = 16


1.2
1
0.8
0.6
0.4
0.2
0
Dispersion variance
0
5
10
//
∞
Block side /km
Point support
Aggregate support


1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
Correlation
0
5
10
15
20
25
Standard deviation /metres
Point support
G-BASE aggregate support
NSI aggregate support