How should a spatial-coverage sample design for a geostatistical soil survey be supplemented to support estimation of spatial covariance parameters?

R.M. Lark[1] *, B.P. Marchant

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

Abstract

We use an expression for the error variance of geostatistical predictions, which includes the effect of uncertainty in the spatial covariance parameters, to examine the performance of sample designs in which a proportion of the total number of observations are distributed according to a spatial coverage design, and the remaining observations are added at supplementary close locations. This expression has been used in previous studies on numerical optimization of spatial sampling, the objective of this study was to use it to discover simple rules of thumb for practical geostatistical sampling. Results for a range of sample sizes and contrasting properties of the underlying random variables show that there is an improvement on adding just a few sample points and close pairs, and a rather slower increase in the prediction error variance as the proportion of sample points allocated in this way is increased above 10 to 20% of the total sample size. One may therefore propose a rule of thumb that, for a fixed sample size, 90% of sample sites are distributed according to a spatial coverage design, and 10% are then added at short distances from sites in the larger subset to support estimation of spatial covariance parameters.

Keywords. Spatial sampling; Prediction variance; Geostatistics

[1]Now at: *School of Biosciences, University of Nottingham, Sutton Bonington Campus, Sutton Bonington, Leicestershire, LE12 5RD, U.K.*

**E-mail address*: murray.lark@nottingham.ac.uk (R.M. Lark).

1. Introduction

1.1 The problem and its motivation

How should we sample a variable in space to allow geostatistical prediction for an information system or mapping project? This is an important question for the application of geostatistics in soil science, particularly when limited resources are available to support soil sampling in the field and the analysis of sampled material in the laboratory. It is important because the sampling determines both the cost of the survey and the quality of the resulting predictions.

One of the first approaches to this question was made by McBratney et al. (1981) who showed that if the spatial covariance parameters (variogram parameters) of the target random variable are known, at least approximately or from a homologous setting, then one may identify the spacing of a square sample grid such that the kriging variance at the centre of a grid cell (where the point kriging variance takes its largest value) does not exceed some threshold. Van Groenigen et al. (1999) demonstrated that spatial simulated annealing, a method for numerical optimization, can be used to find sampling designs in irregularly-shaped regions so as to minimize the mean or maximum kriging variance over that region. This approach will tend to produce a 'space-filling' or 'spatial-coverage' design, which can also be achieved by the methods of Walvoort et al (2010).

The limitation of spatial-coverage designs for geostatistics, be these regular grids or space-filling designs in irregular regions, is that they do not provide information on spatial dependence over short intervals, and so the modelled spatial covariance at short lag distances is poorly constrained. The covariance at short distances is particularly influential on the kriging weights. While some early geostatistical studies in soil science used regular sampling grids (e.g. Burgess and Webster, 1980; Webster and Oliver, 1989) it was realized that it is necessary to include some observations within a sample array that are a short distance apart to support the estimation of spatial covariance parameters (e.g. Atteia et al., 1994; Cattle et al., 2002). However, we are not aware of an explicit analysis

of the benefits of doing this in terms of the quality of final kriging predictions. Stein (1999), in a simple 1-D simulation with only 20 sample locations on a regular transect, showed that the likelihood function for spatial covariance parameters was very flat near the maximum, but that adding just three additional observations at finer intervals within the transect markedly reduced the uncertainty. In 2-D simulations with more realistic sample sizes Haskard (2007) supported this finding. She considered a total sample size of 100, but allocated either 10 or 20 of these points to clusters within an incomplete 10×10 square grid. She found a marked reduction in the standard errors of spatial covariance parameters when using the sample array with 10 points in a cluster by comparison to the full 10×10 grid, and only a small additional benefit in using 20 of the 100 points in clusters.

Simple spatial-coverage sampling will not do to support geostatistical prediction, so how can appropriate designs be discovered? Zhu and Stein (2006) and Marchant and Lark (2007a,b) showed how to define an overall objective function for the quality of a sampling design, an expected mean square error of predictions, which accounts for the two sources of uncertainty in the empirical best-linear unbiased prediction (E-BLUP, equivalent to the kriging prediction in the general case with no covariates and the local mean assumed to be stationary). These two sources are the spatial variation of the target variable and the uncertainty in the maximum likelihood (ML) estimates of the spatial covariance parameters. More detail is provided in section 1.2. The key point is that we do not assume that the spatial covariance parameters are known without error, but account for their uncertainty, which depends in part on the sampling design. Spatial simulated annealing can then be used to minimize the mean value, or the maximum value, of this objective function across a study area. The resulting designs resemble a spatial-coverage sample with some additional points at shorter distances.

These formal methods for optimization may be complex to implement. They require an approximation of the spatial covariance parameters of the target variable, or a specifi-

cation of their joint prior distribution. In practice the scientist who is planning a survey may have a more-or-less fixed sample size to deploy, and simple rules of thumb may be more useful than complex procedures for optimization, which may also be computationally demanding. There are various rules of thumb in geostatistics which have been influential amongst practitioners. For example, it is generally advised to form empirical estimates of the variogram for lag distances no longer than $D/2$ where D is the maximum distance between observations (Journel and Huijbregts, 1978). Webster and Oliver (1992) suggest that at least 100 observations are required to obtain a reliable estimate of the variogram. Kerry et al. (2010) advise that a sampling grid for geostatistical prediction should have a spacing no coarser than half the range of spatial dependence of the target variable, and ideally one third to two fifths of the range.

The objective of this paper is to see whether it is possible to devise rules of thumb to plan a geostatistical soil survey *de novo*. Following the observations of Stein (1999) and Haskard (2007), and from the simulation results of Zhu and Stein (2006) and Marchant and Lark (2007a,b), we propose that the rule for a geostatistical survey with N observations is to withhold some number of these (a short-distance subset), distribute the remaining N_{SC} according to a spatial-coverage design and then to insert each observation from the short-distance subset into the resulting regular array at some fixed short distance, but in a random direction, from a randomly selected site in the spatial-coverage subset. We examine a quality measure for the resulting surveys, the mean square error of prediction as computed by Marchant and Lark (2007a) which accounts both for the density of sampling around a prediction site and the uncertainty of the spatial variance parameters. The key question is whether a general recommendation can be made as to how many sample sites to reserve for the short-distance subset. Our study is therefore one in the spirit of 'innovization' (innovation by optimization), as discussed by Deb et al. (2014). The key idea of innovization is that one seeks to discover rules which a practitioner can implement which capture the key properties of solutions identified by formal optimization.

In the next section we review the calculation of the prediction error variance of the E-BLUP as proposed by Zhu and Stein (2006) and Marchant and Lark (2007a). The Methods section then sets out the sampling schemes and scenarios for which we evaluated this error variance. The scenarios correspond to random variables with a range of spatial covariance parameters. These include parameter sets selected from a Markov Chain Monte Carlo sample of parameters for the random effects in a linear mixed model for the variation of soil carbon content across a part of eastern lowland England with a range of contrasting land uses.

1.2 The mean-square error of the empirical best linear unbiased prediction

In this paper we consider the case of ordinary kriging, although the formulation of the problem extends to the more general best-linear unbiased prediction (BLUP) which includes universal kriging (or regression kriging in an approximately equivalent presentation). The ordinary kriging prediction of a variable, Z at a location $\mathbf{x}_0$, given q covariance parameters in $\boldsymbol{\theta}$ and n observations in $\mathbf{z} = (z(\mathbf{x}_1), z(\mathbf{x}_2), \ldots, z(\mathbf{x}_n))^{\mathrm{T}}$ can be written as

$$\widehat{Z}(\mathbf{x}_0|\boldsymbol{\theta}) \;=\; \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{z}, \tag{1}$$

where $\boldsymbol{\lambda}$ is a vector of weights. The weights are obtained from the ordinary kriging equation

$$\mathbf{L} \;=\; \mathbf{A}^{-1}\mathbf{b}, \tag{2}$$

where

$$\mathbf{A} \;=\; \begin{bmatrix} \mathbf{C}, & \mathbf{1}_n \\ \mathbf{1}_n^{\mathrm{T}}, & 0 \end{bmatrix}$$

the matrix $\mathbf{C}$ is the covariance matrix of the n observations given their locations, $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$ and the covariance function with parameters in $\boldsymbol{\theta}$, $C(\mathbf{x}_i - \mathbf{x}_j|\boldsymbol{\theta})$; $\mathbf{1}_n$ is a vector length n of ones,

$$\mathbf{L} \;=\; \begin{bmatrix} \boldsymbol{\lambda} \\ \psi \end{bmatrix},$$

where ψ is a Lagrange multiplier. If $\mathbf{c}$ is a vector of the covariances between the target location $\mathbf{x}_0$ and the observations, $\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_n$, then

$$\mathbf{b} \;=\; \begin{bmatrix} \mathbf{c} \\ 1 \end{bmatrix}.$$

In this formulation the expected square error of the prediction, the kriging variance, is

$$\sigma^2_{\mathrm{OK}}(\mathbf{x}_0) \;=\; C\left(0|\boldsymbol{\theta}\right) - \mathbf{L}^{\mathrm{T}}\mathbf{b}. \tag{3}$$

The derivations above are based on known covariance parameters, $\boldsymbol{\theta}$. In this paper we consider a frequentist framework in which $\boldsymbol{\theta}$ is treated as fixed but unknown, with the estimate $\widehat{\boldsymbol{\theta}}$ obtained by maximum likelihood, see Lark (2000) for a fuller account. The estimate $\widehat{\boldsymbol{\theta}}$ is 'plugged in' to the equations above to give the empirical BLUP (E-BLUP). Zimmerman & Cressie (1992) considered the effect of this parameter uncertainty on the kriging prediction using a Taylor series approximation. They showed that the prediction remained approximately unbiased, but an additional component of the prediction error variance should be considered. This is

$$\begin{aligned}\tau^2(\mathbf{x}_0) \;&=\; \mathrm{E}\left[\left\{\widehat{Z}\left(\mathbf{x}_0|\boldsymbol{\theta}\right) - \widehat{Z}\left(\mathbf{x}_0|\widehat{\boldsymbol{\theta}}\right)\right\}^2\right] \\ &=\; \sum_{i=1}^{q}\sum_{j=1}^{q}\mathrm{Cov}\left(\theta_i,\theta_j\right)\frac{\partial\widehat{Z}}{\partial\theta_i}\frac{\partial\widehat{Z}}{\partial\theta_j},\end{aligned} \tag{4}$$

where θ_i denotes the ith parameter in $\boldsymbol{\theta}$ and $\mathrm{Cov}\left(\cdot,\cdot\right)$ denotes the covariance of two random terms in the brackets. Zhu and Stein (2006) and Marchant and Lark (2007a) used this as a basis for a component of the expected squared prediction error. The expected value of the term due to uncertainty in the ML estimate, $\widehat{\boldsymbol{\theta}}$, is

$$\mathrm{E}\left[\tau^2(\mathbf{x}_0)\right] \;=\; \sum_{i=1}^{q}\sum_{j=1}^{q}\mathrm{Cov}\left(\theta_i,\theta_j\right)\frac{\partial\boldsymbol{\lambda}^{\mathrm{T}}}{\partial\theta_i}\mathbf{C}\frac{\partial\boldsymbol{\lambda}}{\partial\theta_j}, \tag{5}$$

where $\mathbf{C}$ is the covariance matrix of the n observations, given their locations. The term $\frac{\partial\boldsymbol{\lambda}}{\partial\theta_j}$ is the vector of partial derivatives of the kriging weights with respect to the jth covariance parameter in $\boldsymbol{\theta}$. Marchant and Lark (2007a) provided the following equation from which these can be obtained:

$$\frac{\partial\mathbf{L}}{\partial\theta_i} \;=\; \mathbf{A}^{-1}\left(\frac{\partial\mathbf{b}}{\partial\theta_i} - \frac{\partial\mathbf{A}}{\partial\theta_i}\mathbf{A}^{-1}\mathbf{b}\right). \tag{6}$$

The covariance matrix for the estimated variance parameters may be approximated by the inverse of the Fisher information matrix, $\mathbf{F}$, so

$$\mathrm{Cov}\left(\theta_i,\theta_j\right) \;\approx\; \mathbf{F}^{-1}\left(\theta_i,\theta_j\right), \tag{7}$$

where

$$\mathbf{F} \;=\; \frac{1}{2}\mathrm{Tr}\left[\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \theta_i}\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \theta_j}\right], \tag{8}$$

and $\mathrm{Tr}\left[\cdot\right]$ denotes the trace of the matrix in the brackets (Kitanidis, 1987).

Following Zimmermann and Cressie (1992) we may obtain an overall mean square error of the prediction at $\mathbf{x}_0$, $\sigma^2_{\mathrm{P}}(\mathbf{x}_0)$ as the sum of the kriging variance, Eq. (3), and the expected value of $\tau^2(\mathbf{x}_0)$ given in Eq. (5):

$$\sigma^2_{\mathrm{P}}(\mathbf{x}_0) \;=\; \sigma^2_{\mathrm{OK}} + \mathrm{E}\left[\tau^2(\mathbf{x}_0)\right]. \tag{9}$$

It is acknowledged that this expression is an approximation, given the Taylor Series approximation in Eq. (4) and the comparable assumption in the use of the Fisher Information matrix to obtain $\mathrm{Cov}\left(\theta_i, \theta_j\right)$ in Eq. (7). However, Zhu and Stein (2006) suggest that this approximation is reasonable, at least for comparison between sampling designs.

In this paper we use Eq. (9) to compute the mean squared error of E-BLUPs from samples in which a specified proportion of all observations are distributed according to a spatial-coverage design with each of the remaining points added to a location a short fixed distance from a randomly selected point in the spatial-coverage subset. By varying the total sample size, and the numbers of points in the spatial-coverage subset we were able to show how the division of total sampling effort between spatial-coverage and close points affects the uncertainty in the predictions, and how this differs between random variables with contrasting spatial covariance parameters.

3. Materials and Methods

3.1. Sampling schemes and their implementation.

We start with a fixed total sample size, N. Of these N points $N_{\mathrm{SC}} < N$ were distributed according to a spatial-coverage design within a square uniform region. In the initial experiment N was set to 100 and the uniform region was 256×256 units. The selection of locations for the spatial-coverage points was done with the stratify procedure

in the spcosa library for the R platform (Walvoort et al., 2010; R Core Team, 2014). This procedure uses a k-means algorithm to partition a region into k units, the centroids of which constitute a spatial-coverage sample, Walvoort et al (2010) give further details. The remaining $N - N_{\mathrm{SC}}$ points were then each allocated at random to one of the points in the spatial-coverage subsample and placed a fixed distance ($\delta = 5$ units) from the allocated point in a random direction. The fixed distance, δ, is the 'short' distance included in the sampling scheme to support spatial covariance modelling. We specified a fixed distance (in a random direction) for simplicity. Figure 1 shows the mean distance between a location in the region and its nearest neighbouring sample point in the spatial coverage design. The distance between a location and its nearest neighbouring sample point is the shortest lag over which the spatial covariance is required to determine the E-BLUP prediction. The value of δ was set to a short distance relative to the values in Figure 1, about one fifth the mean distance for the denser sample schemes.

Having generated this sample the next objective is to estimate the maximum prediction error variance at unsampled locations, which is equivalent to finding the kriging variance at the centre of a regular grid cell in the procedure of McBratney et al. (1981). To do this we first found the Voronoi tesselation of the spatial-coverage sample points (the short-distance subset was excluded) using the deldir package in R (Turner, 2015). We then found for each vertex of the set of Voronoi polygons the longest distance to a spatial-coverage sample point in one of its adjacent polygons and then found the maximum value of this distance over all vertices within the central 150×150 unit region, denoted d_{max}. We then found a vertex, at location $\mathbf{x}_{\mathrm{Vmax}}$ and a spatial-coverage point in one of the adjoining polygons $\mathbf{x}_{\mathrm{SCmax}}$ such that the vector $\mathbf{d} = \mathbf{x}_{\mathrm{SCmax}} - \mathbf{x}_{\mathrm{Vmax}}$ has Euclidean norm $|\mathbf{d}| = d_{\mathrm{max}}$. The vertex at $\mathbf{x}_{\mathrm{Vmax}}$ is necessarily a point in the domain such that none other is further from any point in the spatial-coverage set, and so is a site where the kriging variance for a prediction from points in the spatial-coverage set is large.

We then selected five target locations at which the prediction error variance, σ^2_{P},

was evaluated. The first location was at $\mathbf{x}_{\mathrm{Vmax}}$, the fifth location was at $\mathbf{x}_{\mathrm{Vmax}} - \mathbf{d}(|\mathbf{d}| + (\delta/2))/|\mathbf{d}|$. This latter location is on a line joining $\mathbf{x}_{\mathrm{Vmax}}$ to $\mathbf{x}_{\mathrm{SCmax}}$ and is distance $\delta/2$ from $\mathbf{x}_{\mathrm{SCmax}}$. The remaining three points were spaced equally on the line joining the first and fifth. The contribution to the prediction error variance from the kriging variance will be largest at the target point coincident with $\mathbf{x}_{\mathrm{Vmax}}$ and will be smallest at the point closest to $\mathbf{x}_{\mathrm{SCmax}}$. We computed the prediction error variance for a random variable at these five locations, given a set of spatial covariance parameters for the random variable.

We computed the Fisher Information Matrix, Eq. (7), for the spatial covariance parameters using all N observations, but assumed that only the N_{SC} points were available for prediction. This allows us to consider the 'worst case' scenario, i.e. prediction in a region where only points from the spatial-coverage sample are close by. The partial derivatives with respect to the κ and ϕ parameters of the Matérn spatial covariance model (Stein, 1999) were estimated numerically using the grad function from the numDeriv package for the R platform (Gilbert and Varadhan, 2015). This was done because, as discussed by Haskard et al. (2007), the analytical solutions to these derivatives are prone to rounding error. The expected value of τ^2 was then computed at each of the five target points using Eq. (5) and (6). The corresponding kriging variances were computed with Eq. (3) and then the overall prediction error variance at each site was computed with Eq. (9). The maximum prediction error variance over the five sites was then extracted.

The maximum prediction error variance obtained this way for a given sample size, N, with N_{SC} distributed by spatial-coverage sampling is a random variable because the spatial-coverage sample obtained may differ from one run of the stratify procedure to another. For this reason we repeated this procedure 50 times and calculated the mean and 95% confidence interval for each N and N_{SC}.

3.2. Scenarios. 3.2.1 Contrasting random variables. In all calculations we considered a standard random variable with a nugget variance of 0.1 and a correlated variance of 0.9.

We specified an isotropic Matérn spatial correlation model (Matérn, 1986; Stein, 1999)

$$\rho(\mathbf{h}) \;=\; \left\{2^{\kappa-1}\Gamma(\kappa)\right\}^{-1}\left(\frac{|\mathbf{h}|}{\phi}\right)^{\kappa} K_{\kappa}\left(\frac{|\mathbf{h}|}{\phi}\right), \tag{10}$$

where $\mathbf{h}$ is the lag vector, $\Gamma(\cdot)$ is the gamma function, K_κ is a modified Bessel function of the second kind of order κ. The two parameters are κ, a smoothness parameter, and ϕ, a distance parameter. We considered cases with three values of κ: 0.5 (equivalent to the exponential variogram), 0.3 (somewhat rougher than exponential) and 0.7 (somewhat smoother than exponential). All three smoothness parameters were specified in combination of each of four effective ranges of spatial dependence at which the correlation decays to approximately 0.05, equal to $\alpha_\kappa\phi$ where ϕ is the distance parameter of the variogram and α_κ is equal to 2.4, 3.0 and 3.5 respectively when κ is 0.3, 0.5 and 0.7. We specified values of ϕ such that the effective range was 37.5, 50, 75 or 150 units. We do not claim that this range of scenarios is exhaustive. It reflects a case with a small nugget effect and with the smoothness of the random process in the vicinity of that of an exponentially correlated one. Within this region of the feature space of spatial covariance parameters we consider the effects of variations in the effective range and the overall sample size.

For any given set of spatial covariance parameters we considered three sample sizes with N= 75, 100 and 150. In each case we considered spatial-coverage subsets with $N_{\mathrm{SC}} = N - n$ with $n = 1, 2, \cdots, 10$ and then larger numbers which differed between the basic sample sizes. Figure 1 shows how the mean distance from a point in the spatial-coverage sample to its nearest neighbour depends on the number of points in the spatial-coverage sample.

We kept the overall sampling density constant, irrespective of N. To achieve this the dimensions of the uniform square region of interest were adjusted so that the overall sampling density was equal to $100/256^2$ in all cases. This means that differences between sample sizes are not confounded with effects of sample density. The effect of sample density can be examined by comparing cases with fixed sample size and different effective ranges of spatial dependence.

3.2.2 Random variables in linear mixed models for variation of soil properties . Here we used data from a previously-published study on spatial variation of soil at within-field scale in a lowland landscape in eastern England (Lark et al., 1998). The data were collected from a 6-ha field in use for research on precision agriculture management of cereal crops. The soil was sampled in the spring in the presence of a winter-sown barley crop but prior to any fertilizer application. The soil was sampled at 100 locations which comprised a basic 50-m grid (24 points) with 66 additional sample points at 10-m intervals on transects aligned with the 50-m grid, and 10 sample points added to allow comparisons over 5 m. At each site the soil was sampled to depth 20 cm with a screw auger. In this paper we use data on the organic matter content of the soil samples, which was determined by the Walkley Black method (Hesse, 1971), and the nitrate content extractable by KCl (Mengel, 1991). Statistical analysis was conducted on the organic matter content, expressed as a percent by mass of the (dry) soil, and on the natural logarithm of nitrate content (mg kg^{-1} dry soil).

In both cases a linear mixed model was fitted to the data with a constant mean as the only fixed effect. We fitted a Matérn correlation function for the correlated random effect; the parameters of this function, along with a nugget variance and the variance of the spatially correlated random effects, were estimated by ML.

We considered a situation in which a 16-ha square site, considered to be homologous with the original field with respect to soil variation, is to be sampled to allow geostatistical mapping of both variables. We assume that the total sample size is fixed at 100. We followed the same procedures described in section 3.1 to generate realizations of space-filling samples with between 99 and 50 sample points, and with the remaining sample sites (1 to 50) distributed between sites of the space-filling design selected at random, and placed 5 m from the associated site in a random direction. For each design we computed the maximum prediction error variance in the same way described in section 3.1.

4. Results

Results are shown in Figures 2 to 5. In Figure 2 are shown results for all cases where the effective range of the random variable of interest was 150 units. This is large relative to the spacings between points in the spatial-coverage samples as shown in Figure 1. Figure 2a and 2b both show the maximum prediction error variance for random variables with the parameter κ equal to 0.7 (somewhat smoother than an exponential random variable). The different symbols correspond to the different sample sizes, and the horizontal bars show the 95% confidence interval for each mean value. In Figure 2a the maximum prediction error variance is plotted against the fraction of all N sample sites which are in the short-distance subset (rather than the spatial-coverage subset) up to a maximum proportion of 0.4. In Figure 2b the same results are shown but plotted against the number of sample points in the short-distance subset. Similarly, Figures 2c and 2d show maximum prediction error variance plotted against, respectively, the proportion of sites in the short-distance subset and the number for a random variable with $\kappa = 0.5$, and Figures 2e and 2f are corresponding plots for the case with $\kappa = 0.3$. Figures 3, 4 and 5 show the corresponding output for cases with the effective range of the random variable equal to 75, 50 and 37.5 units. The latter is of similar size to the distance to the nearest point in the spatial-coverage samples.

In Figure 2 in all cases there is an initial reduction in the maximum prediction error variance as a result of increasing the number of sample sites in the short-distance set from one and a more gradual increase in the maximum prediction error variance as the number of short-distance sites is increased much above 10 or so. Reducing κ (making the random variable rougher) makes the response of the maximum prediction error variance to the number of short-distance sample points more sensitive. For the random variable with $\kappa = 0.3$ it can be seen that there is no improvement from increasing the number of short-range sites above about six, but the curves are somewhat 'flat-bottomed', and there is very little increase in the maximum prediction error variance if up to about 10 points are allocated to the short-range subset. Figure 2e shows that increasing the proportion of

points in the short-distance set above about 0.1 increases the maximum prediction error variance for all sample sizes. For the variables with κ equal to 0.5 or 0.7 (Figures 2a–2d) the 'flat-bottomed' form of the plots is more pronounced, with very little increase in the maximum prediction error variance as the size of the short-distance subset is increased.

Reducing the effective range of the random variable, other factors being held constant, increases the maximum prediction error variance, as can be seen by comparing Figure 2 with Figures 3–5. It also makes the increase in the maximum prediction error variance as the short-distance subset is increased above 10–15 points more pronounced, and the effect is very notable for the random variable with the shortest effective range (Figure 5), although this also depends on the total sample size. Note that the range of values on the ordinate of the plots (prediction error variance) is increased for shorter effective ranges, and that a wider range is used for Figure 3e and f than for the plots for the smoother random variables with an effective range of 75 units.

In the case of the random variables with effective range of 75 or 50 units the maximum prediction error variance is markedly reduced on adding up to 2 points in the short-range subset when κ is equal to 0.5 or 0.7, but for the rougher random variable with $\kappa = 0.3$ further improvement is achieved by adding 5 to 7 points in the short-range set. Using up to 10 sample points in the short-range set incurs a small penalty for the smoothest random function ($\kappa = 0.7$) with the smallest sample size, but the increase in the maximum squared prediction error is not large, and with a total sample size of 100 or 150 the increase is negligible for up to 20 or so points in the short-distance set.

Figure 5 shows results for the case where the effective range is 37.5, short relative to the spacing between neighbouring sites in the spatial-coverage sample. Note that in many cases the maximum squared prediction error exceed the *a priori* variance of the random variable. Whilst there is a benefit from putting some points into the short-distance set the increase in prediction error variance from putting too many into this set is very pronounced for the smaller two sample sizes. Reduction in the spatial-coverage subset of points, with

the addition of extra points at a short distance, affects the uncertainty in the spatial covariance parameters as well as the kriging variance component of the prediction error variance when the range of spatial dependence is close to the spacing of the spatial-coverage subset.

Comparing Figure 2e and 2f show that the absolute number of short-distance points rather than the proportion of points in the subset determines the initial reduction in the maximum prediction error variance (a short range set of 5-7 achieves the minimum squared-prediction error regardless of the overall sample size. However, the increase in the maximum prediction error variance with the proportion of sample points in the short-range set is similar for all sample sizes as this proportion increases above 0.1. Examining all the plots shows that setting the number of short-distance points to 10% of the total sample size (shown by the vertical dotted line in the plots for the short-distance fraction) ensures that sufficient short-distance points are included. For the random functions equivalent to the exponential or rougher the curves are sufficiently flat that a 10% rule incurs no penalty from reducing the spacing of the spatial-coverage sampling, and any such effect for the smoothest random variable considered is very small.

Figure 6 shows the scaled variograms for organic matter content and for nitrate content from the sampled field. In each case the value of the variogram is divided by the sill to facilitate comparison. Note that while the values are similar over longer lags the behaviour at short lags is rather different. In the case of nitrate content there is a large nugget effect, but the parameter κ for the correlated random variable is 1.04, implying a random process which is smoother than an exponentially correlated one. In the case of organic matter content the parameter κ is 0.12, which implies a markedly rougher process. The nugget effect in this latter case is zero.

Figure 7 shows the variogram models, scaled to an *a priori* (sill) variance of 1, for organic matter and the log of nitrate content, as estimated by ML. The parameters are listed in Table 1. Figure 8 shows the maximum prediction error variance for different

numbers of sample points out of 100 used for short-distance comparisons. The prediction error variances are standardized by the *a priori* variance of the respective random effects. In both cases there is a marked reduction in the error variance on the adjustment of the sample design to include some observations at short distance. The minimum error variance for organic matter predictions is with 8 observations used to allow comparisons over short distances, and for nitrate content the minimum is with 17 such observations. In both cases the rate of increase in the prediction error variance when more observations than the optimal number are used for short-distance comparisons is markedly less than the rate of increase as fewer such observations are included. That said, the reduction in the prediction error variance for nitrate as more than 10 or so observations are used is very small. If one was planning a survey to map both these variables, a design with 90 observations distributed for spatial coverage and 10 included subsequently at short distances, would not be markedly suboptimal for either variable. It is interesting that the 10% rule, which seemed reasonably robust for the hypothetical examples with the κ parameter close to 0.5 and a nugget effect equal to one tenth of the sill, is also reasonable for these two soil variables with rather different values of both parameters.

5. Discussion and Conclusions

These results show that the findings of Stein (1999) and Haskard (2007) that a relatively small subset of short-distance points in a sample can markedly improve the estimation of the covariance parameters extends to the corollary that these short-distance points can also improve the maximum prediction error variance, which reflects the uncertainty in both the covariance parameters and the spatial variation between target points for prediction and their neighbouring observations. Figure 1 shows that the distance between nearest neighbours in a spatial coverage sample increases relatively slowly as the sample density is reduced over the range considered in this study. This, with the findings of Stein (1999) and Haskard (2007) account for the asymmetry of the plots in Figures 2–5 with a reduction in the prediction error variance on the initial addition of a few close

points which is steeper than subsequent increases in the prediction error variance as the number of points in the spatial coverage sample declines.

The practical conclusion is that it is important to include a short-distance subset. With the larger sample size considered here, and particularly for random variables as rough or rougher than the exponential, the potential cost of under-investing in short-distance sampling is larger than the cost from degrading the spatial-coverage set by including sample points in the short-distance subset, at least as long as the spatial-coverage set is not too coarse relative to the effective range of the random variable.

While a very small number of sample points may markedly improve the maximum prediction error variance in these examples, it must be recalled that these calculations are done on the assumption of second-order stationarity. If just two or three short-distance points are included then there is a risk that they will appear in atypical conditions, and this could have a substantial effect on the estimated covariance parameters. For this reason the inclusion of a rather larger short-distance set is good practice, and these results suggest that using about 10% of the total sample effort in a short-distance subset is reasonable.

In this study we restrict the supplementation of the spatial coverage sample to single points at a fixed distance from one of the spatial coverage set. In the sample schemes presented by Marchant and Lark (2007a) that optimize a mean prediction error variance the outputs resemble spatial coverage samples with either close pairs or, for some sets of covariance parameters, short 'transects'. This suggests that there might be scope for further studies in the spirit of innovization to uncover rules which relate the supplementation strategy to the properties of the underlying random variable. Whether this provides a basis for practical rules of thumb would depend on, first, the sensitivity of the optimal strategy to the (unknown) properties of the underlying random variable and, second, how far the robustness of this strategy depends on the stationarity assumption. Additional questions for further work would include whether it is more effective to include supplementary points at a fixed distance from the spatial coverage points, or to include

them at random distances bounded by a maximum.

The expression for the prediction error variance used in this paper is for the case of the E-BLUP where the mean is treated as an unknown constant, equivalent to ordinary kriging. There is no reason why this should not be ezxtended to a more general case where the mean is modelled as a function of some environmental covariates. We would not expect the general conclusions to differ much, because supplementary points will have a negligible effect on estimation of fixed effects coefficients, but an investigation of the question would be a useful further study.

To conclude, on the basis of these results we may recommend that, provided the spatial-coverage subset of sample points is sufficiently dense to be reasonably confident that the spatial dependence of the target variable is resolved, it is good practice to include a short-distance subset and a relatively small investment of sample effort in such sample points, which add little to the field effort required for sampling, can have a large effect on the uncertainty of kriging predictions. In our hypothetical examples, where the nugget effect is small to moderate (10%) and the smoothness parameter is close to that for the exponential variogram (with both rougher and smoother conditions considered) a robust strategy is to use a short-distance subset that corresponds to about 10% of the total sample size for a range of values of the effective range of spatial dependence. We showed how the optimal size of the short-distance subset could be found using the variogram for two variables in a real data set on the soil. Interestingly, applying the 10% rule would be a robust strategy for both these variables although their variograms are rather different from those used in the hypothetical examples.

Acknowledgements

This paper is published with the permission of the Executive Director of the British Geological Survey (NERC).

References

Atteia, O., Webster, R., Dubois, J.-P. 1994. Geostatistical analysis of soil contamination in the Swiss Jura. Environmental Pollution 86, 315–327.

Burgess, T.M., Webster, R. 1980. Optimal interpolation and isarithmic mapping of soil properties. 1. The semi-variogram and punctual kriging. Journal of Soil Science 31, 315 –331.

Cattle, J.A., McBratney, A.B., Minasny, B. 2002. Kriging method evaluation for assessing the spatial distribution of urban soil lead concentration, Journal of Environmental Quality 31, 1576–1588.

Deb, K., Bandaru, S., Greiner, D., Gaspar-Cunha, A., Tutum, C.C. 2014. An integrated approach to automated innovization for discovering useful design principles: case studies from engineering. Applied Soft Computing. 15, 42–56.

Gilbert, P., Varadhan, R. 2015. numDeriv: Accurate Numerical Derivatives. R package version 2014.2-1. http://CRAN.R-project.org/package=numDeriv

Haskard, K.A., 2007. An anisotropic Matérn spatial covariance model: REML estimation and properties. PhD Thesis, University of Adelaide, South Australia. Available at http://hdl.handle.net/2440/47972

Haskard, K.A., Cullis, B.R., Verbyla, A.P., 2007. Anisotropic Matérn correlation and spatial prediction using REML. Journal of Agricultural, Biological and Environmental Statistics 12, 1–14.

Hesse, P. R. 1971. A Textbook of Soil Chemical Analysis. London, John Murray.

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

Kerry, R., Oliver, M.A., Frogbrook, Z.L. 2010. Sampling in precision agriculture. In: Geostatistical Applications for Precision Agriculture (ed M.A. Oliver), pp. 35–63. Springer, Dordrecht.

Kitanidis, P. K., 1987, Parametric estimation of covariances of regionalised variables. Water Resources Bulletin, 23, 557–567.

Lark, R.M. 2000. Estimating 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., 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.

Marchant, B.P., Lark, R.M. 2007a. Optimized sample schemes for geostatistical surveys. Mathematical Geology, 39, 113–134.

Marchant, B.P., Lark, R.M. 2007b. The Matérn variogram model: implications for uncertainty propagation and sampling in geostatistical surveys. Geoderma, 140, 337–345.

Matérn, B. 1986. Spatial Variation, Lecture Notes in Statistics, No. 36, Springer, New York.

McBratney, A.B., Webster, R., Burgess, T.M. 1981. The design of optimal sampling schemes for local estimation and mapping of regionalised variables. I. Theory and Method. Computers and Geosciences 7, 331–334.

Mengel, K., 1991. Available nitrogen in soils and its determination by the Nmin method and by electroultrafiltration (EUF). Fertilizer Research, 12, 37-52.

R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Stein, M.L., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York.

Turner, R. 2015. deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tessellation. R package version 0.1-9. http://CRAN.R-project.org/package=deldir

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

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

Webster, R., Oliver, M.A. 1989. Optimal interpolation and isarithmic mapping of soil properties. VI. Disjunctive kriging and mapping the conditional probability. Journal of Soil Science, 40, 497–512.

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

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

Zhu, Z., Stein, M.L. 2006. Spatial sampling design for prediction with estimated parameters. Journal of Agricultural, Biological and Environmental Statistics, 11, 24–44.

Zimmerman, D.L., Cressie, N. 1992. Mean squared prediction error in the spatial linear model with estimated covariance parameters. Annals of the Institute of Statistical Mathematics, 44, 27–43.

Table 1. Parameters of the variograms estimated by maximum likelihood for soil organic matter content and (log-transformed) nitrate content.

Variable	ϕ /m	κ	Nugget variance	Correlated variance
Organic matter /%	97.9	0.12	0.00	0.92
Nitrate /log mg kg^{-1}	34.9	1.04	0.14	0.15

Figure captions

Figure 1. Mean distance to nearest neighbour within a set of N_{SC} points in a spatial coverage sample in a 256 × 256-unit square region.

Figure 2. Maximum prediction error variance over the standard set of five prediction locations as described in the text for random variables with an effective range of 150 units and $\kappa = 0.7$ (Fig. 2a,b), 0.5 (Fig. 2c,d) or 0.3 (Fig 2e,f). The plotted value is the mean over 50 realizations of the sampling scheme with horizontal bars showing the 95% confidence interval. The total sample size is indicated by the plotted symbol 150 (▲), 100 (○) or 75 (●). In the left-hand column (Fig 2a, 2c and 2e) the mean value of the maximum prediction error is plotted against the proportion of the total sample size witheld from the spatial coverage subset and added as the short-distance subset (up to a maximum proportion of 0.4). In the right-hand column (Fig. 2b, 2d and 2f) the same values are plotted against the number of points in the short-distance subset.

Figure 3. Maximum prediction error variance over the standard set of five prediction locations as described in the text for random variables with an effective range of 75 units and $\kappa = 0.7$ (Fig. 3a,b), 0.5 (Fig. 3c,d) or 0.3 (Fig 3e,f). The plotted value is the mean over 50 realizations of the sampling scheme with horizontal bars showing the 95% confidence interval. The total sample size is indicated by the plotted symbol 150 (▲), 100 (○) or 75 (●). In the left-hand column (Fig 3a, 3c and 3e) the mean value of the maximum prediction error is plotted against the proportion of the total sample size witheld from the spatial coverage subset and added as the short-distance subset (up to a maximum proportion of 0.4). In the right-hand column (Fig. 3b, 3d and 3f) the same values are plotted against the number of points in the short-distance subset.

Figure 4. Maximum prediction error variance over the standard set of five prediction

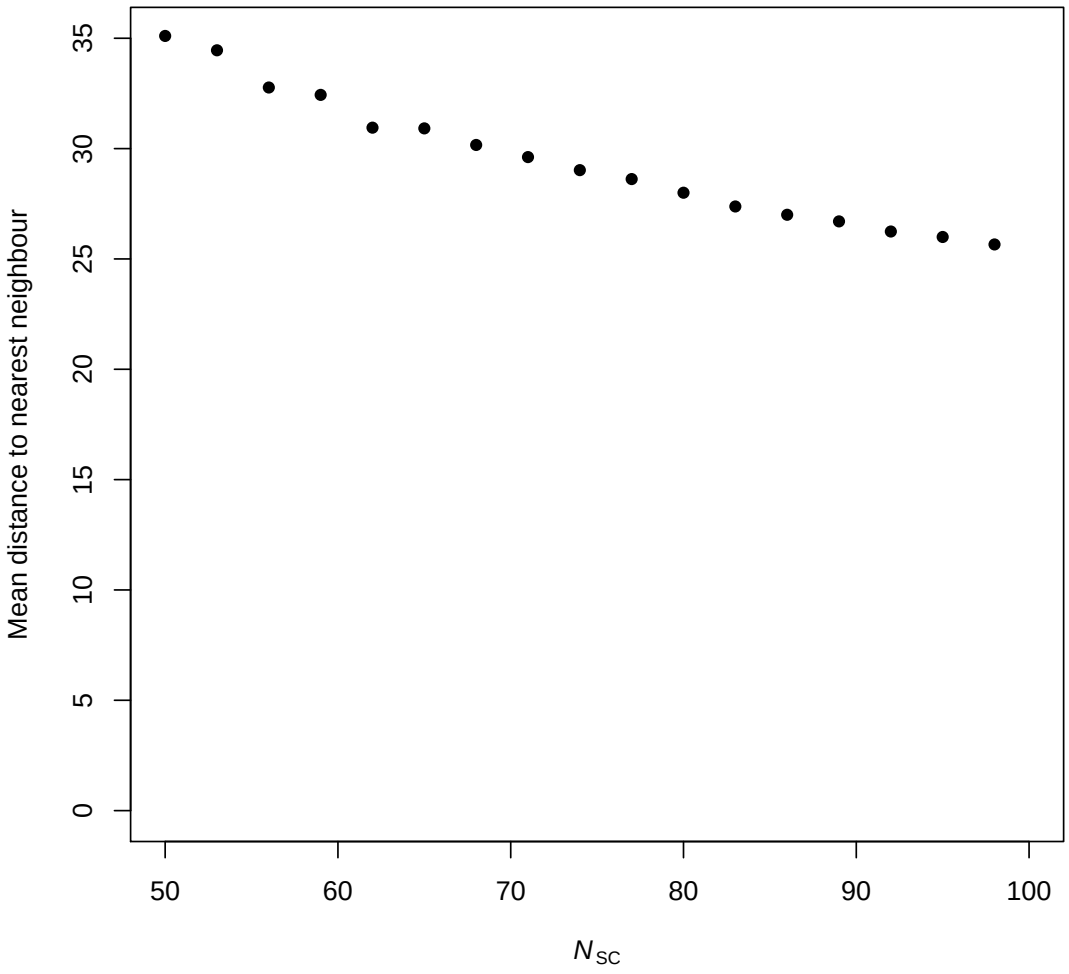
locations as described in the text for random variables with an effective range of 50 units and $\kappa = 0.7$ (Fig. 4a,b), 0.5 (Fig. 4c,d) or 0.3 (Fig 4e,f). The plotted value is the mean over 50 realizations of the sampling scheme with horizontal bars showing the 95% confidence interval. The total sample size is indicated by the plotted symbol 150 (▲), 100 (○) or 75 (●). In the left-hand column (Fig 4a, 4c and 4e) the mean value of the maximum prediction error is plotted against the proportion of the total sample size witheld from the spatial coverage subset and added as the short-distance subset (up to a maximum proportion of 0.4). In the right-hand column (Fig. 4b, 4d and 4f) the same values are plotted against the number of points in the short-distance subset.

Figure 5. Maximum prediction error variance over the standard set of five prediction locations as described in the text for random variables with an effective range of 37.5 units and $\kappa = 0.7$ (Fig. 5a,b), 0.5 (Fig. 5c,d) or 0.3 (Fig 5e,f). The plotted value is the mean over 50 realizations of the sampling scheme with horizontal bars showing the 95% confidence interval. The total sample size is indicated by the plotted symbol 150 (▲), 100 (○) or 75 (●). In the left-hand column (Fig 5a, 5c and 5e) the mean value of the maximum prediction error is plotted against the proportion of the total sample size witheld from the spatial coverage subset and added as the short-distance subset (up to a maximum proportion of 0.4). In the right-hand column (Fig. 5b, 5d and 5f) the same values are plotted against the number of points in the short-distance subset.

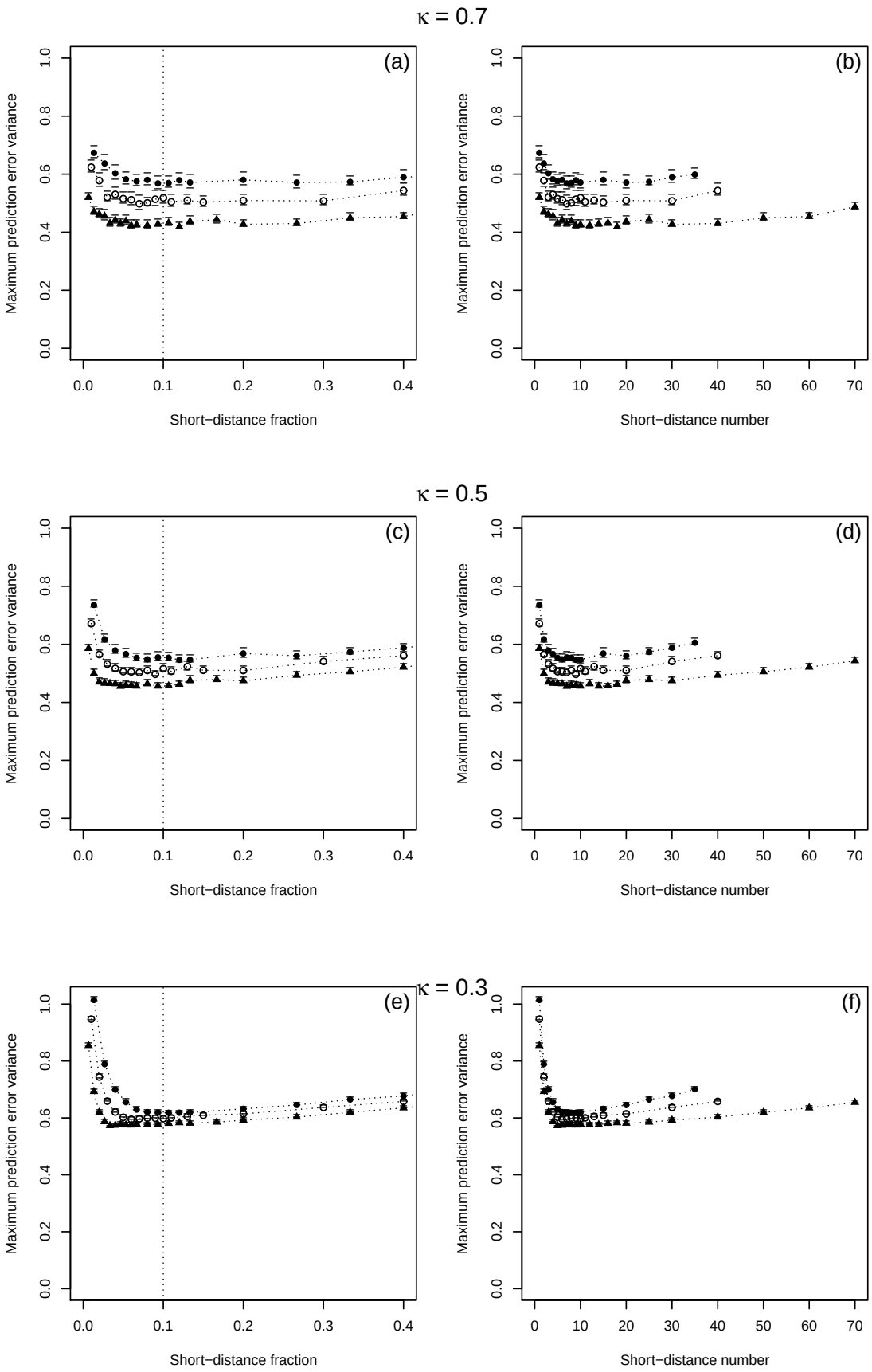
Figure 6. Variograms for organic matter content (broken line) and log of nitrate content (solid line) estimated by maximum likelihood and scaled to *a priori* (sill) variance of 1.

Figure 7. Mean maximum prediction error variance for organic matter content (●) and log of nitrate content (○) with a total sample size of 100 in a 16-ha square region. The number of sample sites inserted at short distances in a space-filling design varies

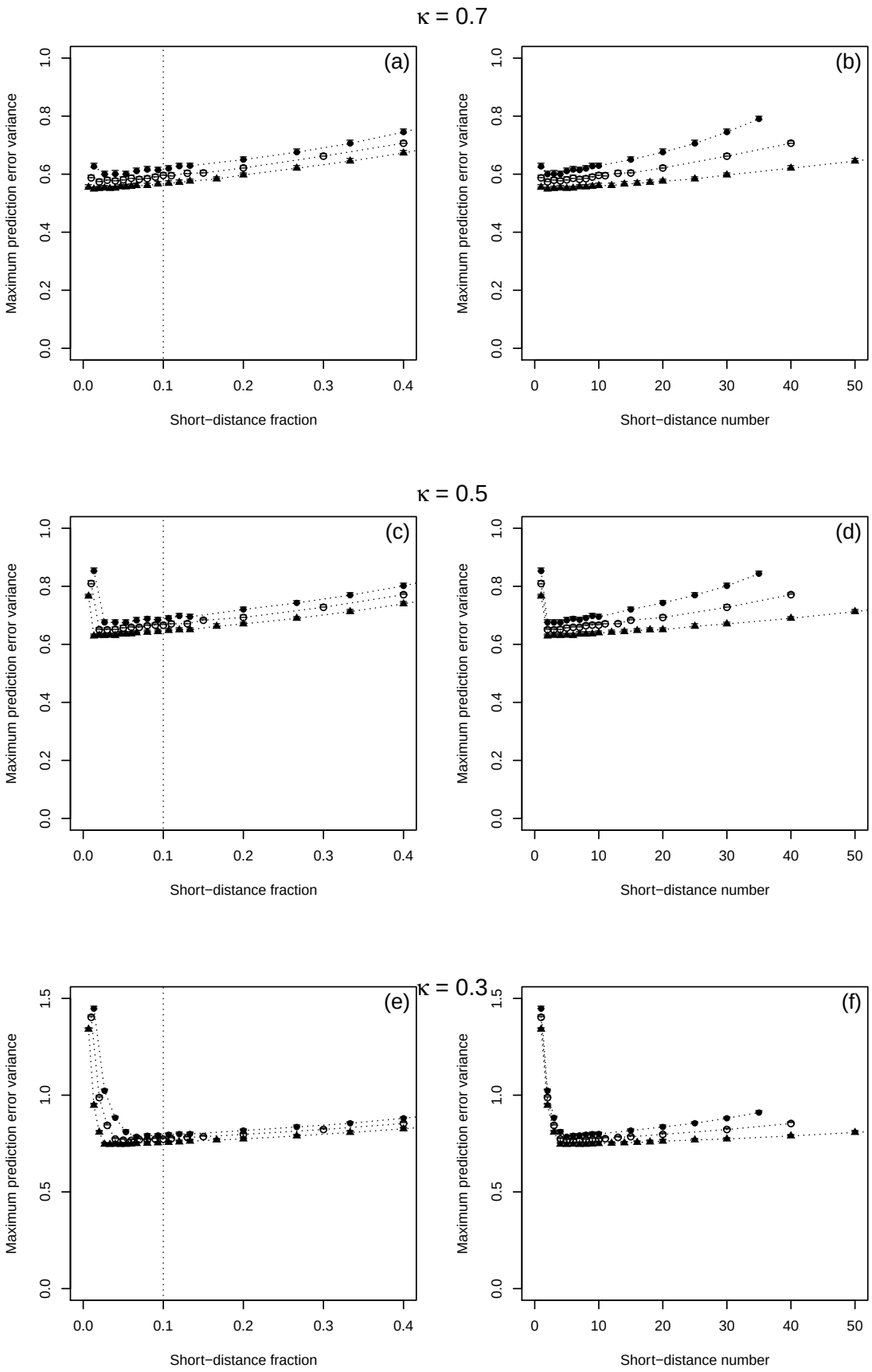
from 1 to 50. Crosses show the 95% confidence interval. The dotted vertical line shows the design where the mean maximum prediction error variance is smallest for organic matter content, and the dashed line shows the same for log-transformed nitrate content.


Mean distance to nearest neighbour
0
5
10
15
20
25
30
35
50
60
70
80
90
100
N_{SC}

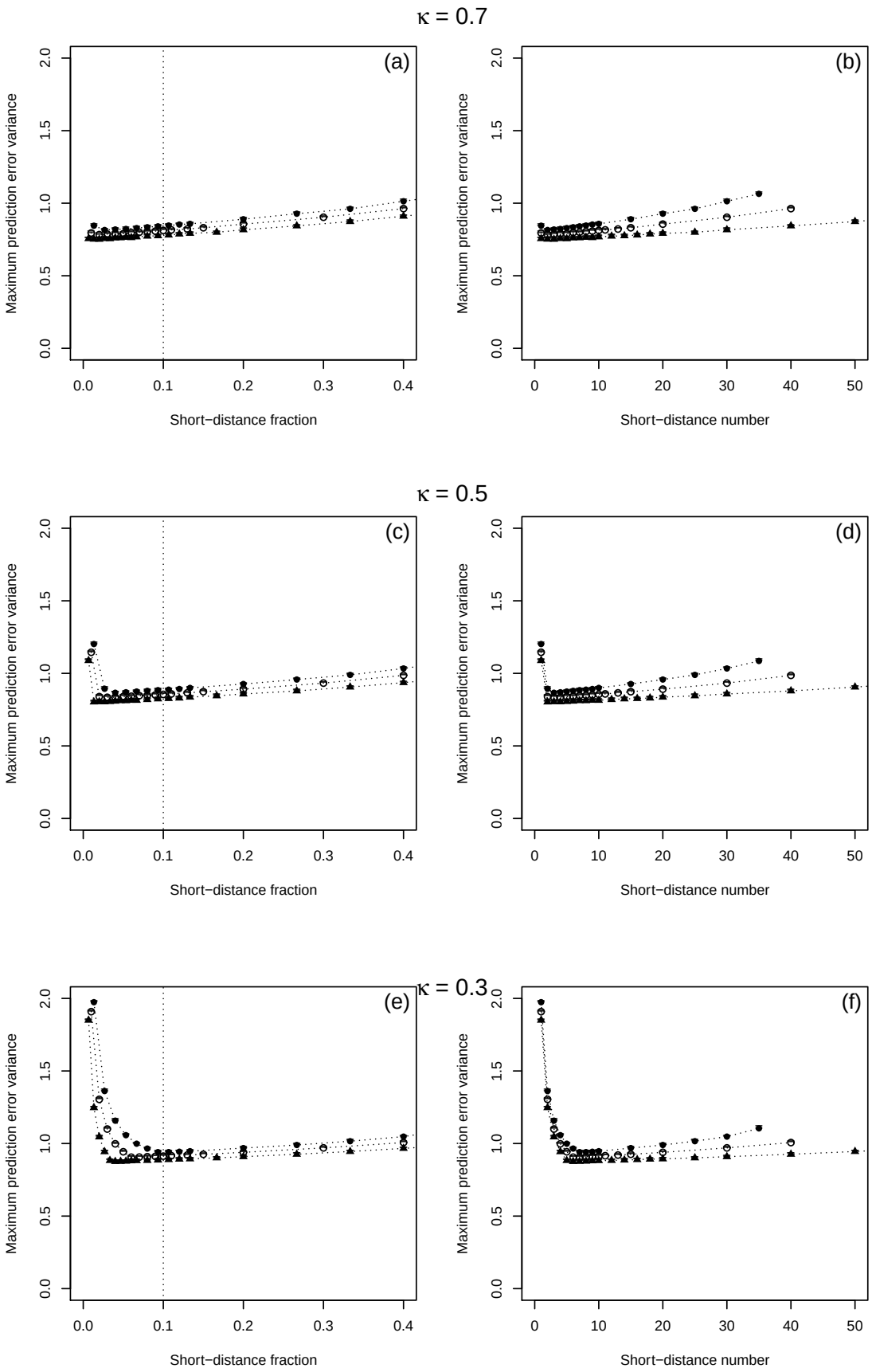
1: Fig 1


Effective range 150
κ = 0.7
(a)
(b)
κ = 0.5
(c)
(d)
κ = 0.3
(e)
(f)
Maximum prediction error variance
Short−distance fraction
Short−distance number
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
70

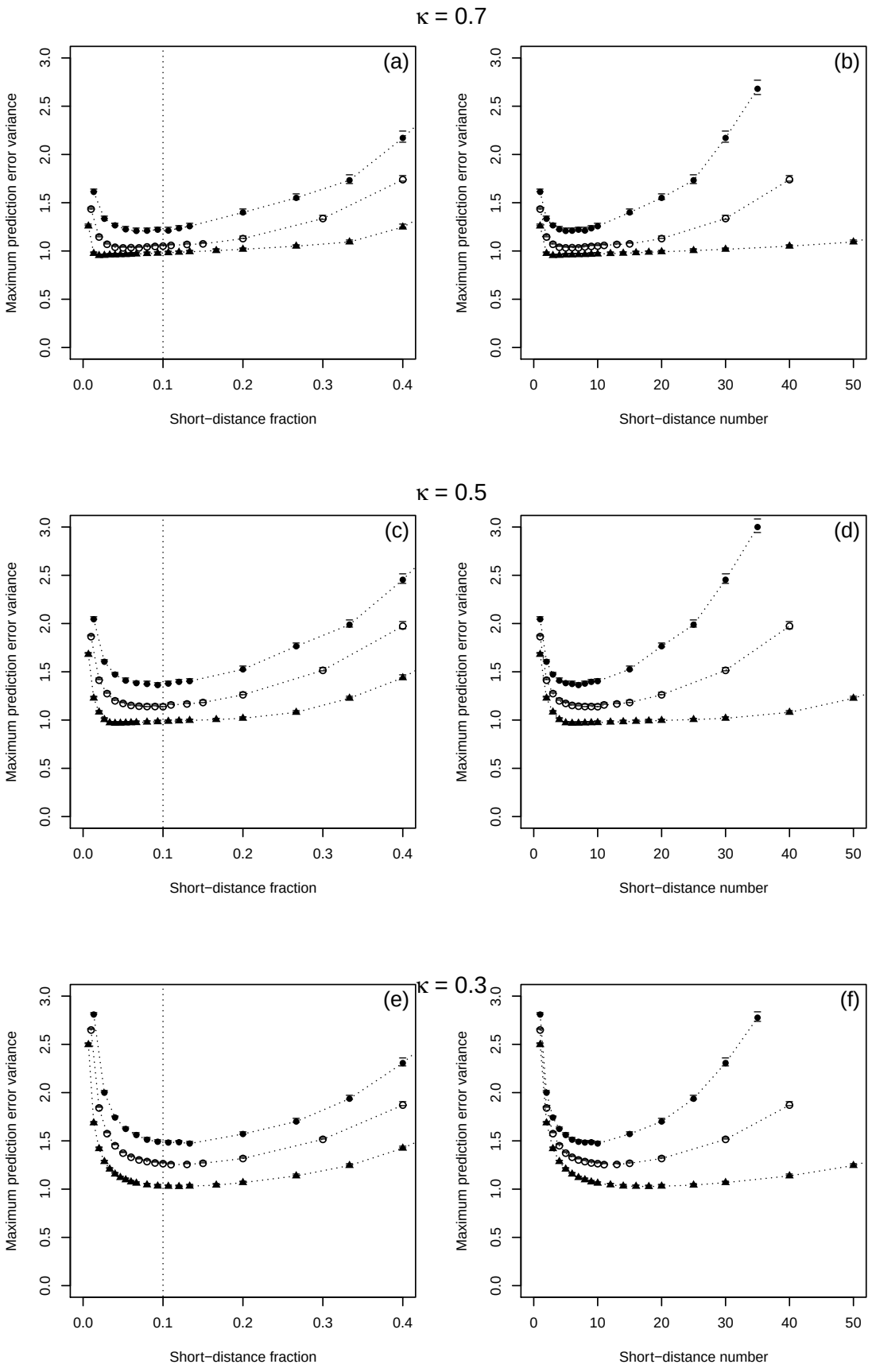
2: Fig 2


Effective range 75
κ = 0.7
(a)
(b)
κ = 0.5
(c)
(d)
κ = 0.3
(e)
(f)
Maximum prediction error variance
Short−distance fraction
Short−distance number

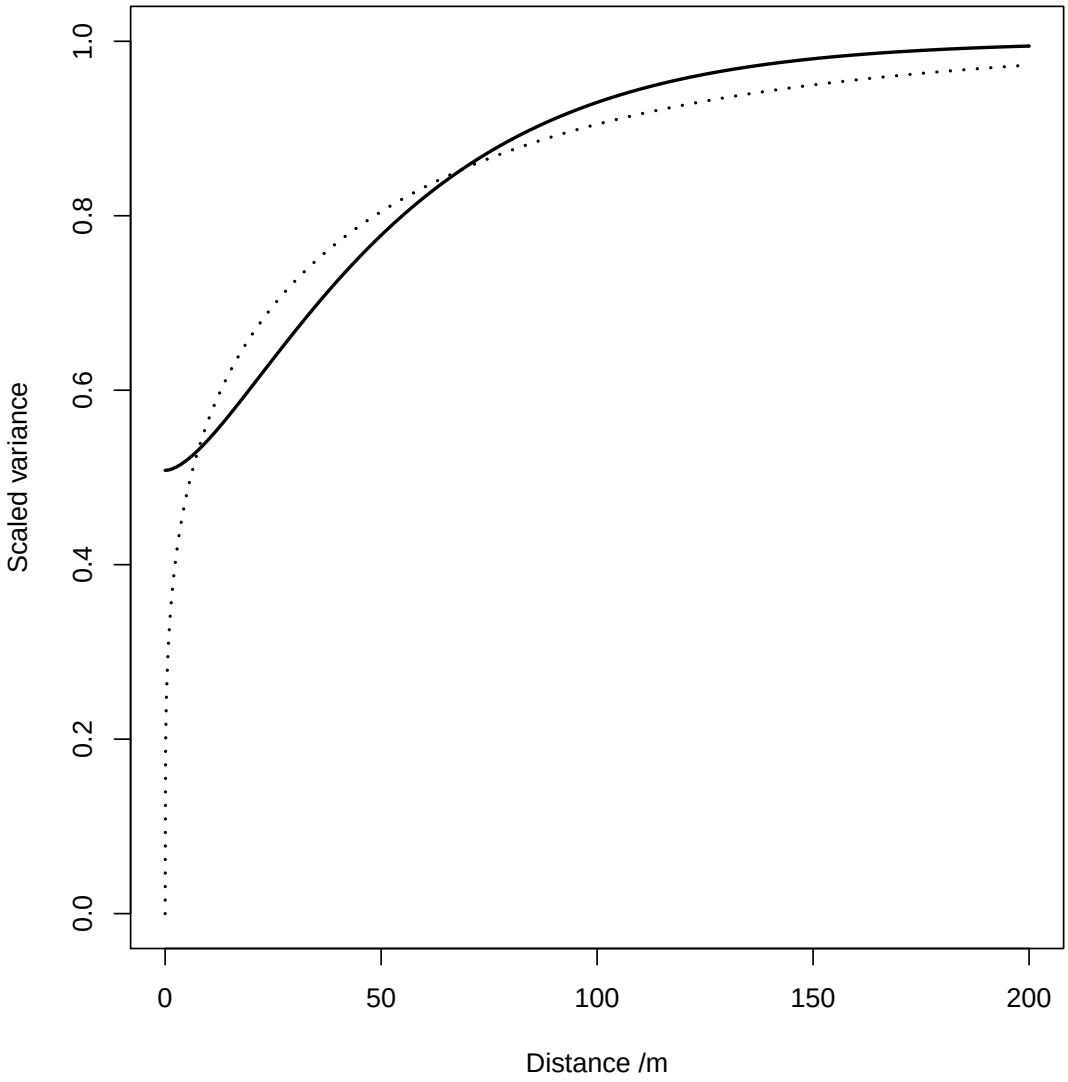
3: Fig 3


Effective range 50
κ = 0.7
(a)
(b)
κ = 0.5
(c)
(d)
κ = 0.3
(e)
(f)
Maximum prediction error variance
Short−distance fraction
Short−distance number
0.0
0.5
1.0
1.5
2.0
0.1
0.2
0.3
0.4
10
20
30
40
50

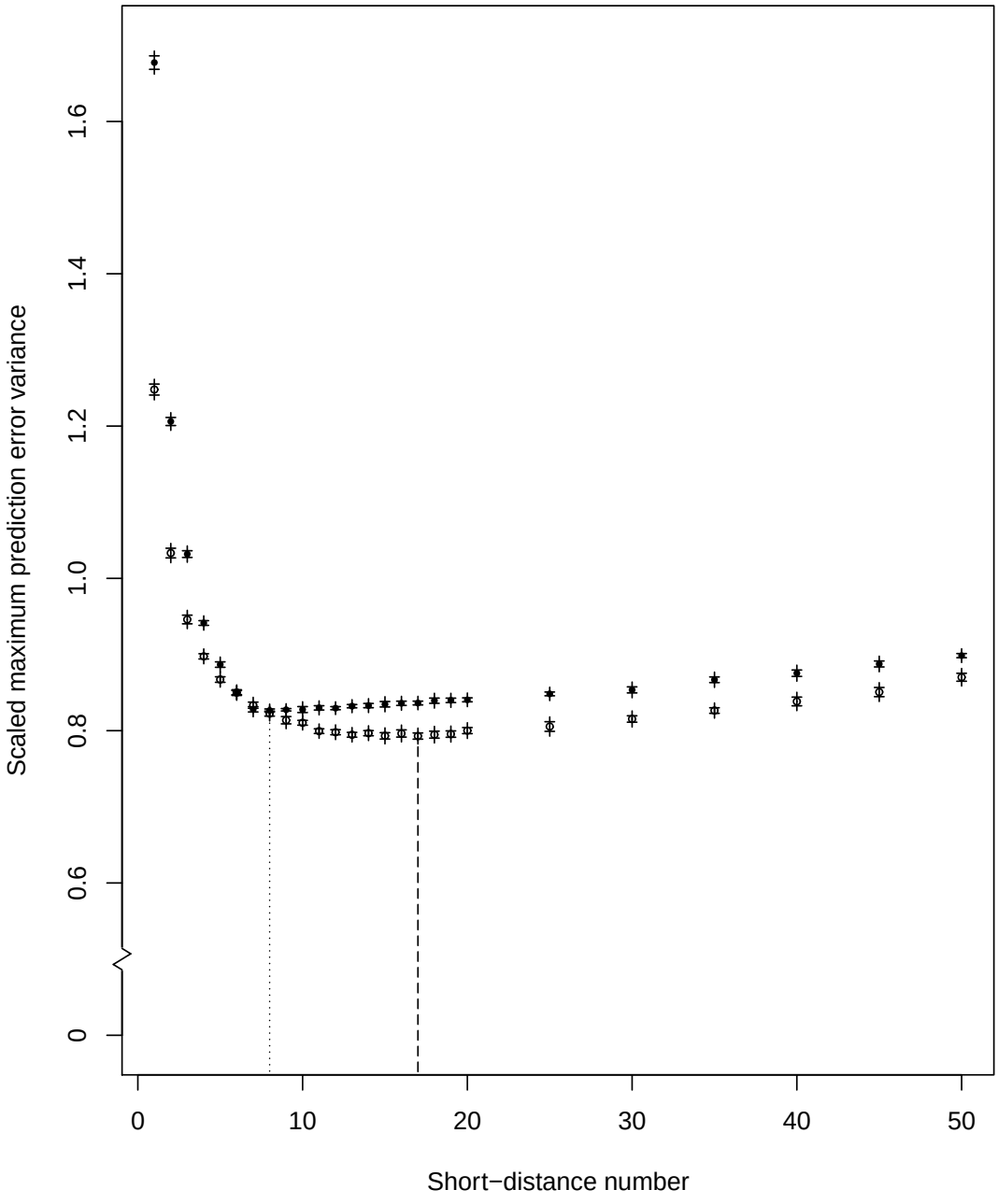
4: Fig 4


Effective range 37.5
κ = 0.7
(a)
(b)
κ = 0.5
(c)
(d)
κ = 0.3
(e)
(f)
Maximum prediction error variance
Short−distance fraction
Short−distance number
0.0
0.5
1.0
1.5
2.0
2.5
3.0
0.1
0.2
0.3
0.4
10
20
30
40
50

5: Fig 5



6: Fig 6


Scaled maximum prediction error variance
0
0.6
0.8
1.0
1.2
1.4
1.6
0
10
20
30
40
50
Short−distance number

7: Fig 7