Using measurements close to a detection limit in a geostatistical case study to predict selenium concentration in topsoil

T.G. Orton [a, *], B.G. Rawlins [b], R.M. Lark [a]

[a] *Rothamsted Research, Harpenden, Hertfordshire. AL5 2JQ, UK*

[b] *British Geological Survey, Keyworth, Nottingham, Notts. NG2 6JJ, UK*

Abstract

Data on environmental variables are subject to measurement error (ME), and it is important that this ME should be considered in any statistical analysis. Environmental datasets commonly consist of positive random variables that have skewed distributions. Measurements are then usually reported with a theoretical detection limit (DL); measurements less than this DL are deemed not to be statistically different from zero, and these data are then treated by setting them to an arbitrary value of half of the DL. The skew of the data is dealt with by taking logarithms, and the geostatistical analysis performed for the transformed variable. The DL approach, however, is somewhat *ad hoc*, and in this paper we investigate an alternative approach to incorporate such measurements in a geostatistical analysis, namely Bayesian hierarchical modelling. This approach incorporates 'soft' data (i.e. imprecise information), and we use soft data to represent the information that each measurement provides. We can use this approach to combine a lognormal model to describe the spatial variability with a Gaussian model for the measurement error. We apply the methods to a dataset on the selenium (Se) concentration in the topsoil throughout the East Anglia region of the UK. We compare the maps of predictions produced by the

[*] Corresponding author.
E-mail address: thomas.orton@bbsrc.ac.uk (T.G. Orton).

approaches, and compare the methods based on their ability to predict the Se concentration and the associated uncertainty. We also consider how the geostatistical predictions might be used to aid the effective management of Se-deficient soils, and compare the methods based on the costs that might be incurred from the selected management strategies. We found that the Bayesian approach based on soft data resulted in smoother maps, reduced the errors of the predictions, and provided a better representation of the associated uncertainty. The cost resulting from Se-deficient soils was generally lower when we used the soft data approach, and we conclude that this provides a more effective and interpretable model for the data in this case study, and possibly for other environmental datasets with measurements close to a DL.

Keywords - Geostatistics, Bayesian hierarchical modelling, measurement error, detection limits.

1. Introduction

Any data collected on a variable are subject to measurement error (ME). In many cases this error is negligible relative to other sources of variation and may effectively be ignored when it comes to manipulating the data set and using it to make predictions of that variable at unsampled locations. However, in other cases, the measurement process can give rise to considerable errors — it is then imperative that this error is considered in any subsequent analysis.

Environmental datasets commonly consist of measurements of non-negative random variables. For example, the concentration of a substance in the soil can obviously not be negative. When taking measurements of such positive random variables, it is usual to report the measurements along with a theoretical detection limit. This detection limit (DL) is a property of the measurement process. It is calculated (from repeated measurements in a control experiment) so that any measurement that is less than this limit is deemed not to be statistically different from zero.

It is common practice in analysis of environmental data with a DL to treat large values as precise (without error) and to set values below the DL to an arbitrary value of half the DL (Woodside and Kocurek, 1997). This approach, however, is somewhat *ad hoc*, and in this paper, we investigate an alternative approach to represent and incorporate data on variables with a DL in a geostatistical analysis.

Classical geostatistics provides a number of approaches by which we may incorporate ME in the analysis of a Gaussian spatial random field (SRF). If the ME is Gaussian with an unknown (but constant) variance, then the nugget effect of the variogram accounts for the error, in which case the nugget effect, c_0, is the sum of two components, the microscale process, c_{MS}, and the measurement error, c_{ME}:

$$c_0 = c_{\mathrm{MS}} + c_{\mathrm{ME}}. \quad (1)$$

In practice, we can only separate the nugget variance in this way if we have duplicate measurements at some locations, so that the measurement error can be estimated. As duplicate measurements are rarely undertaken, the two components are generally unresolved in practical studies. If the measurement error is Gaussian with a known (estimated) variance, then kriging with measurement error (Webster and Oliver, 2007) may be used to calculate predictions. If the measurement error variance, c_{ME}, is unknown, but repeated measurements are available, then we can use the repeated measurements to estimate c_{ME} and incorporate this estimate in the kriging predictions (Laslett and McBratney, 1990).

These methods to incorporate measurement error are based on a Gaussian model for the spatial random field (SRF). However, in this paper, we are concerned with the analysis of positive random variables using data that are subject to ME. Such positive random variables often exhibit strong positive skewness (i.e. many measurements close to zero, and fewer larger measurements, as is often the case for numerous major and trace element concentrations in soils and sediments), in which case the Gaussian assumption for the SRF, *Z*, may not be justified. This skewness can often be removed by taking logarithms (Webster and Oliver, 2007), in which case, a variogram may be fitted for the log-transformed variable, $Y = \ln Z$; in this case, we refer to the original SRF, *Z*, as a lognormal SRF. In the absence of measurement error, kriging may then be used to predict the log-transformed variable, and the prediction transformed back to predict the original variable; if we require that the predictor be unbiased, then the method is called ordinary lognormal kriging (OLK).

If we consider that the nugget of the variogram for a log-transformed SRF, *Y*, incorporates measurement error, as in Eq. (1), then we essentially assume that the

measurement error model is Gaussian for the log-transformed variable, *Y*. Although the Gaussian assumption for the microscale variation of *Y* may be appropriate, it does not follow that the measurement process should also give rise to errors that follow the same pattern, since the generation and the measurement of the SRF are essentially independent processes. The classical measurement error model is Gaussian for the variable that is being measured (i.e. *Z*). For unbiased measurements with a constant measurement error variance, this choice may be justified by the maximum entropy principle (Kapur and Kesavan, 1992). As far as we know, a kriging system for incorporating Gaussian measurement error for *Z* for a lognormal SRF has not been described. The approach that has been commonly adopted to incorporate measurement error in the analysis of lognormal SRFs is the detection limit (DL) approach (Woodside and Kocurek, 1997).

Reimann and Filzmoser (2000) looked at a wide range of variables from environmental datasets, and found that most of these variables exhibited variation that could not be explained by either the normal or the lognormal distribution, but rather originated from more than one process. We can consider a Bayesian approach (Banerjee et al., 2004) to combine a Gaussian measurement error model with a lognormal SRF model, and by doing so, provide one possible approach to deal with such data. The Bayesian approach consists of a prior and a posterior stage. In the prior stage, we choose appropriate probability distributions that represent our beliefs about the values of the parameters (of the mean and covariance models) and variables (i.e. measurable quantities) in the system prior to collecting the data. In the posterior stage, we update these prior beliefs in the light of the data, through Bayesian conditioning; this results in a joint posterior distribution for the variables and parameters, which we may use to make our inferences about the quantities of interest. The Bayesian

approach allows for the inclusion of *soft* data in a spatial analysis; these soft data represent imprecise information, as opposed to the precise measurements that are represented by the *hard* data. Here, we consider how we might use these soft data to represent measurement error. The Bayesian approach also incorporates parameter uncertainty, which can be a considerable advantage in the analysis of lognormal SRFs, where predictions can be sensitive to the fitted variogram parameters. Previous studies that have investigated a Bayesian approach to incorporate imprecise measurements in a spatial analysis include De Oliveira (2005) and Fridley and Dixon (2007). In particular, these studies show how we might incorporate censored measurements; in this work, we aim to incorporate measurements that are subject to measurement error, but from which we can still extract information other than just some censoring limits.

In this paper, we begin by introducing a case study, on the concentration of selenium (Se) in the soil, and demonstrate how we might deal with measurement error in the analysis of a lognormal SRF. We review two approaches to spatial prediction — the classical kriging approach (ordinary and lognormal), and hierarchical Bayesian modelling — paying special attention to how ME is dealt with through these methodologies. We apply the prediction methods to the case study, and discuss the results.

2. Introduction to the case study

Although Se is toxic in excess, it is also an essential element for human health. Low dietary intakes of Se are associated with health disorders, including oxidative stress-related conditions, reduced fertility and immune function, and an increased risk

of cancers (Fan et al., 2008). The amount of Se in the soil is therefore important because it can influence root uptake and crop Se concentration (Adams et al., 2002). The British Geological Survey collected soil samples at 5761 locations throughout the East Anglia region of the UK. The concentration of Se (as well as other elements) in each of these samples was determined using X-ray fluorescence spectrometry (XRF-S) and reported in mg kg^{-1}; see Lark et al. (2006) for a fuller discussion of the sampling and analytical procedures. The measurements ranged from a minimum of -0.1 to a maximum of 9.5, and are plotted as a histogram in Fig. 1a, and on a classified map of the region in Fig. 2. Clearly, a negative concentration is impossible — readings such as this must be due to measurement errors. If a longer period had been devoted to the analysis of each sample, a lower DL would have been achieved; a decision was taken that the benefit of the lower DL was insufficient in comparison to the extra time required.

The geostatistical methods that we will use in this paper are based on a Gaussian model for the log-transformed variable, $Y = \ln Z$. Fig. 1b shows the marginal distribution for *Y*, the log-transform of the Se data; since this shows *Y* to be roughly Gaussian, it supports the assumption.

The objective in this case study was to use the available measurements to predict the Se concentration at any location in the study region, and hence be able to identify more accurately those areas where there was a risk of Se deficiency in the soil. This information could be used to identify sites where an application of Se to the soil might be beneficial.

3. Representing measurement error

We begin by introducing some notation. In the following, we suppose that we have taken a single measurement of the original (i.e. untransformed) variable, *Z*, at each of the *N* data locations, $\mathbf{x}_{\mathrm{D}} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$. We write this vector of measurements as $\boldsymbol{\zeta}_{\mathrm{D}} = (\zeta_1, \zeta_2, ..., \zeta_N)^{\mathrm{T}}$, and refer to the actual values at these locations (i.e. the values that would have been recorded had there been no imprecision in the measurement process) as $\mathbf{z}_{\mathrm{D}} = (z_1, z_2, ..., z_N)^{\mathrm{T}}$. When we refer to a pdf-type soft datum (we will introduce this concept later in this section), we consider the pdf, $f_{\mathrm{s}.i}(z_i|\zeta_i)$ — or simply $f_{\mathrm{s}.i}(z_i)$ for short — that represents the information that we obtain about the variable, $Z(\mathbf{x}_i)$, when we are given the measurement of this variable, ζ_i. We consider independent measurement errors, and we therefore write

$$f_{\mathrm{s}}(\mathbf{z}) = \prod_{i=1}^{N} f_{\mathrm{s}.i}(z_i).$$

The magnitude of measurement error for Se was investigated by the repeated analyses of three soil certified reference materials (CRM). Each homogenized CRM was repeatedly subsampled and made into 22 pellets; each of these pellets underwent the XRF-S analysis to give a measurement of its Se concentration. Results from the repeated CRM measurements are shown in Table 1.

A common approach to deal with measurement error in such geochemical surveys is to adopt a detection limit (DL) approach (Woodside and Kocurek, 1997). The DL for a particular measurement method is defined as the smallest concentration that is statistically different from zero — here we consider the 5 % level for statistical difference. The measurements shown in Table 1, along with other calibration measurements, were used to determine a DL of 0.2 mg kg^{-1}. When incorporating ME in geostatistical predictions via a DL approach, we assume that large measurements

185 can be considered as accurate. Since any measurements that are less than the DL
186 cannot be considered as statistically different from zero, these data are assumed to
187 take the value of half of this DL. All of these data are then assumed to be
188 measurements of the variable of interest in the study, and the geostatistical predictions
189 that result from this approach incorporate measurement error through the nugget
190 variance.

191 Another approach that may be used to incorporate measurement error in a
192 geostatistical analysis is to use soft data; the hierarchical Bayesian approach allows
193 for the inclusion of this kind of data. A soft datum represents the information that we
194 receive from a measurement at a location about the true underlying value at that same
195 location. To derive the form of this soft datum, we assume a measurement error
196 model. In each of the three cases shown in Table 1, the variance of the repeated
197 measurements was approximately 0.01 ($\hat{\sigma}^2_{\text{me}} = 0.01$), despite the mean of the
198 measurements being different. It appears from these measurements that the classical
199 measurement error model would be a reasonable assumption (i.e. measurements are
200 independent and unbiased). Therefore, for a single measurement, ζ_i, at a location, $\mathbf{x}_i$,
201 we write the measurement error model to give the probability (density) of taking this
202 measurement, if the actual concentration being measured were z_i:

203 $$f_{\text{me}.i}\left(\zeta_i \middle| z_i\right) = N\left(\zeta_i; z_i, \hat{\sigma}^2_{\text{me}}\right), \tag{2}$$

204 (i.e. for each location, $\mathbf{x}_i$, $i = 1,...,N$, ζ_i has a Gaussian distribution with a mean of
205 z_i and a variance of $\hat{\sigma}^2_{\text{me}}$). From this measurement error model, we can use Bayes
206 theorem, to write the *n* soft data pdfs:

207 $$f_{\text{s}.i}\left(z_i \middle| \zeta_i\right) \propto \begin{cases} N\left(z_i; \zeta_i, \hat{\sigma}^2_{\text{me}}\right), & z_i > 0 \\ 0, & \text{otherwise}. \end{cases} \tag{3}$$

This is based on a uniform prior for $z_i > 0$, which we can justify here because any other information that we have about z_i (prior to measurement) is accounted for through the geostatistical method that we will use to process these soft data (see the methods section below).

4. Spatial prediction methods

We will consider three geostatistical approaches to estimate the Se concentration at unsampled locations in our case study: ordinary and ordinary lognormal kriging (OK and OLK), and Bayesian estimation. In this section, we briefly review OLK and Bayesian estimation, noting how they can be used to deal with ME. We pay special attention to the Bayesian estimation method, in which we combine ideas from hierarchical modelling (Banerjee et al., 2004) and the Bayesian maximum entropy method (BME; Christakos, 2000).

In the following, we seek a prediction of the variable, *Z*, at the single prediction location, $\mathbf{x}_0$; we refer to this variable as $Z(\mathbf{x}_0)$ or Z_0, and the values that this variable can take as z_0. We use $\boldsymbol{\Sigma}_D$ to refer to the covariance matrix between the data locations, and $\boldsymbol{\Sigma}_{0,D}$ for the vector containing the covariances between the prediction and data locations.

4.1 Ordinary lognormal kriging

If the data exhibit a strong positive skew, then any variogram that is fitted from the data is sensitive to small changes in the larger data values, because of the large contribution they make to the squared differences. This problem can often be

overcome by considering the logarithmic transform, $Y = \ln(Z)$, of the original variable, *Z*. If the transformed SRF, *Y*, is approximately Gaussian, then we can use ordinary lognormal kriging (OLK) to calculate our prediction.

Suppose that we have estimated the variogram from the transformed data, $\mathbf{y}$, that the local mean for *Y* is constant but unknown, and that we seek a prediction of *Z* at the location $\mathbf{x}_0$. The OLK estimate is calculated so that it is unbiased and it minimizes the mean squared logarithmic error, $\mathrm{E}\left[\left(\ln \hat{Z}_{\mathrm{OLK}}(\mathbf{x}_0) - \ln Z_0\right)^2\right]$. Note that it is not guaranteed to also minimize the mean squared error (MSE) on the original scale. If we write $\hat{Y}_{\mathrm{OK}}(\mathbf{x}_0)$ and $\sigma^2_{\mathrm{OK}}(\mathbf{x}_0)$ for the ordinary kriging estimate and variance for the transformed variable, $Y(\mathbf{x}_0)$, then the OLK prediction for $Z(\mathbf{x}_0)$ is given by (see Journel, 1980):

$$\hat{Z}_{\mathrm{OLK}}(\mathbf{x}_0) = \exp\left\{\hat{Y}_{\mathrm{OK}}(\mathbf{x}_0) + \frac{1}{2}\sigma^2_{\mathrm{OK}}(\mathbf{x}_0) - \psi\right\}. \quad (4)$$

where $\psi = \dfrac{\left(1 - \boldsymbol{\Sigma}_{0,\mathrm{D}}\boldsymbol{\Sigma}_{\mathrm{D}}^{-1}\mathbf{1}\right)}{\mathbf{1}^{\mathrm{T}}\boldsymbol{\Sigma}_{\mathrm{D}}^{-1}\mathbf{1}}$. Note that ψ is often called the Lagrange multiplier — the Lagrange multiplier, however, depends on the precise form of the Lagrangian used in the constrained optimization (which is not unique), and as such we prefer to specify the equation for ψ here.

If we do not require that the predictor be unbiased, and require a predictor that minimizes the error on the logarithmic scale, we can use the median predictor:

$$\tilde{Z}_{\mathrm{OLK}}(\mathbf{x}_0) = \exp\left\{\hat{Y}_{\mathrm{OK}}(\mathbf{x}_0)\right\}, \quad (5)$$

where we use the tilde to denote the median predictor. Note that the back-transform does not depend on the kriging variance here, and so the predictor is not so sensitive

253 to small changes in the variogram parameters. As a result, this predictor is often
254 preferred in the literature (e.g. Tolosana-Delgado and Pawlowsky-Glahn, 2007).
255 To measure the uncertainty, we use a confidence interval; we calculate this
256 confidence interval from the kriging prediction for the log-transformed variable, since
257 we can directly transform the quantiles back to give a confidence interval for the
258 original variable.
259 Ordinary lognormal kriging cannot be used to process pdf-type soft data.
260 Measurement error is dealt with in OLK either through the variogram or by assuming
261 that the data are subject to a detection limit (DL). If it is dealt with solely through the
262 variogram, then the error is assumed to be Gaussian for the transformed variable, *Y*. In
263 this case, the variogram should be fitted from all of the log-transformed data. A DL
264 approach assumes that large measurements can be considered as precise, whilst any
265 measurements less than the DL are imprecise, and given the value of half of this DL.
266 If we use these values to fit the variogram, then we will underestimate the nugget
267 variance, since we will have many identical values in our dataset. We therefore use
268 just the larger values (i.e. those above the DL) to fit the variogram.
269
270 *4.2 Hierarchical Bayesian modelling approach*
271 We split our description of the hierarchical Bayesian modelling approach into
272 three sections: first we describe the model, second we state the prior distributions for
273 the model parameters, and third we show how we can implement the hierarchical
274 Bayesian modelling approach through Markov chain Monte Carlo (MCMC) methods.
275 Banerjee et al. (2004) provide a useful textbook, introducing the theory and
276 application of Bayesian modelling for the analysis of spatial data.
277

1. Model – The hierarchical Bayesian model that we use for our data here can perhaps best be pictured through a directed graph model, as shown in Fig. 3. This essentially consists of four components: the trend and covariance models, the SRF model, the transformation of this SRF, and the measurement error model. The trend and covariance models and transformation of the SRF are deterministic (i.e. given the inputs, the outputs are defined uniquely by the model or transformation). The SRF and measurement error models are stochastic (i.e. given the inputs to these models, the output is a random sample from some probability distribution that is parameterized by the model inputs). We now describe each of these components in turn, starting with the trend and covariance models.

We denote the trend and covariance parameters as the vector, $\boldsymbol{\theta}$. In our case study, we will consider a constant mean, μ, and an exponential covariance model, so that $\boldsymbol{\theta} = \left(\mu, \sigma^2, s, a\right)$, where σ^2 is the total variance, *s* is the proportion of this variance with a spatial structure, and *a* is the effective range of correlation. In Fig. 3, we include the trend parameters, $\beta_0, \beta_1, ..., \beta_p$, which can be used to model a non-constant mean function, although in this work we will consider a constant mean only, so that we have $\boldsymbol{\mu} = \beta_0 \mathbf{1}$, and we write μ in place of β_0. For the covariance matrix, $\boldsymbol{\Sigma}$, we use the exponential model with a nugget effect. We can write the covariance matrix as $\boldsymbol{\Sigma} = \sigma^2 \mathbf{A}$, where $\mathbf{A}$ is the correlation matrix for the data and prediction locations.

Given the mean vector, $\boldsymbol{\mu}$, and covariance matrix, $\boldsymbol{\Sigma}$, the model for the SRF, *Y*, is Gaussian with these parameters. This constitutes the second section of the graph model in Fig. 3. The SRF, *Y*, is for some transform of our original variable, *Z*, and we will come to this transform next.

Since the data in our case study exhibit strong positive skewness, the Gaussian SRF model for the original variable, *Z*, is inappropriate. In such circumstances, a common approach is to assume that some transform of the data, $Y = \varphi(Z)$, gives a variable, *Y*, for which the Gaussian assumption is appropriate. The logarithmic transform is commonly used in geostatistics to perform this task (as in OLK) for positively skewed data. This transformation is shown as the third section in Fig. 3. If we assume that the transformed variable takes a Gaussian distribution, and if we are given the mean and covariance parameters, $\boldsymbol{\theta}$, then we can write the SRF model for the values of *Z* (at the prediction and data locations) as:

$$f_{\text{SRF}}(\mathbf{z}_{0\text{D}}|\boldsymbol{\theta}) = J_{\varphi}\, MVN(\mathbf{y}_{0\text{D}}; \boldsymbol{\mu}(\mu), \boldsymbol{\Sigma}(\sigma^2, s, a)), \tag{7}$$

where $J_{\varphi} = \prod_{i=0}^{N} 1/z_i$ (the Jacobian determinant of the transformation) is the product of the inverses of the elements in the vector, $\mathbf{z}_{0\text{D}} = \begin{bmatrix} z_0 \\ \mathbf{z}_{\text{D}} \end{bmatrix}$, and $MVN(\mathbf{y}_{0\text{D}}; \boldsymbol{\mu}(\mu), \boldsymbol{\Sigma}(\sigma^2, s, a))$ is the multivariate Gaussian model for the transformed variable, $\mathbf{y}_{0\text{D}}$, parameterized by the mean vector, $\boldsymbol{\mu}(\mu)$, and covariance matrix, $\boldsymbol{\Sigma}(\sigma^2, s, a)$. Essentially, this represents the assumption that the transformed SRF, *Y*, is Gaussian, and that *Y* is the log-transform of the variable, *Z*. Eq. (7) then effectively models the first three sections in the graphical model, Fig. 3.

We can include measurement error in a Bayesian hierarchical modelling approach using the measurement error model, $f_{\text{me}}(\boldsymbol{\zeta}_{\text{D}}|\mathbf{z}_{\text{D}})$, (as described in Section 3). This measurement error model comprises the bottom section of Fig. 3, which completes the Bayesian hierarchical model that we consider for our case study.

The graphical model helps us to picture the relationships between the parameters and variables in the system. We must now write our joint statistical model for these parameters and variables:

$$f(\boldsymbol{\theta}, \mathbf{z}_{0\mathrm{D}}, \boldsymbol{\zeta}_{\mathrm{D}}) = f_0(\boldsymbol{\theta})\, f_{\mathrm{SRF}}(\mathbf{z}_{0\mathrm{D}}|\boldsymbol{\theta})\, f_{\mathrm{me}}(\boldsymbol{\zeta}_{\mathrm{D}}|\mathbf{z}_{\mathrm{D}}). \tag{8}$$

This is simply an application of the fundamental rule of probability (i.e. for two events, A and B, $\Pr[A, B] = \Pr[A]\Pr[B|A]$), and the assumption that the measurements, $\boldsymbol{\zeta}_{\mathrm{D}}$, and the parameters, $\boldsymbol{\theta}$, are conditionally independent given the actual values of the SRF, $\mathbf{z}_{0\mathrm{D}}$. Eq. (8) gives the joint probability of observing any combination of values, $\boldsymbol{\theta}, \mathbf{z}_{0\mathrm{D}}, \boldsymbol{\zeta}_{\mathrm{D}}$; we are interested in the probabilities for the values of Z_0, given the measurements, $\boldsymbol{\zeta}_{\mathrm{D}}$. We can calculate these probabilities by integrating out the unknowns here (i.e. the parameters, $\boldsymbol{\theta}$, and the actual values, $\mathbf{z}_{\mathrm{D}}$, of the SRF at the data locations — not the measurements). This gives us our prediction distribution:

$$\begin{aligned} \pi(z_0|\boldsymbol{\zeta}_{\mathrm{D}}) &\propto \int f(\boldsymbol{\theta}, \mathbf{z}_{0\mathrm{D}}, \boldsymbol{\zeta}_{\mathrm{D}})\, \mathrm{d}\mathbf{z}_{\mathrm{D}}\, \mathrm{d}\boldsymbol{\theta} \\ &= \int f_0(\boldsymbol{\theta}) f_{\mathrm{SRF}}(\mathbf{z}_{0\mathrm{D}}|\boldsymbol{\theta}) f_{\mathrm{me}}(\boldsymbol{\zeta}_{\mathrm{D}}|\mathbf{z}_{\mathrm{D}})\, \mathrm{d}\mathbf{z}_{\mathrm{D}}\, \mathrm{d}\boldsymbol{\theta}. \end{aligned} \tag{9}$$

We now define the soft data as the information provided about the SRF, $\mathbf{z}_{\mathrm{D}}$, by the measurements, $\boldsymbol{\zeta}_{\mathrm{D}}$, and we can use Bayes's theorem to write:

$$f_{\mathrm{s}}(\mathbf{z}_{\mathrm{D}}|\boldsymbol{\zeta}_{\mathrm{D}}) = \frac{f_{\mathrm{me}}(\boldsymbol{\zeta}_{\mathrm{D}}|\mathbf{z}_{\mathrm{D}}) f_{\mathrm{z}}(\mathbf{z}_{\mathrm{D}})}{\int_{\mathbf{0}}^{\infty} f_{\mathrm{me}}(\boldsymbol{\zeta}_{\mathrm{D}}|\mathbf{z}_{\mathrm{D}}) f_{\mathrm{z}}(\mathbf{z}_{\mathrm{D}})\, \mathrm{d}\mathbf{z}_{\mathrm{D}}}, \tag{10}$$

where $f_{\mathrm{z}}(\mathbf{z}_{\mathrm{D}})$ is a 'prior' distribution for $\mathbf{z}_{\mathrm{D}}$. The hierarchical Bayesian approach is based on the assumption that the measurements are conditionally independent of the parameters, $\boldsymbol{\theta}$, given the values of the SRF, $\mathbf{z}_{0\mathrm{D}}$; we do not have to account for the spatial correlation in $\mathbf{z}_{0\mathrm{D}}$ through $f_{\mathrm{z}}(\mathbf{z}_{\mathrm{D}})$ here (and therefore choose a uniform prior for $f_{\mathrm{z}}(\mathbf{z}_{\mathrm{D}})$ over the positive numbers), because their spatial correlation is already

accounted for through the SRF model, $f_{\mathrm{SRF}}\left(\mathbf{z}_{0\mathrm{D}}|\boldsymbol{\theta}\right)$. The normalization constant in the denominator of Eq. (10) does not depend on $\mathbf{z}_{\mathrm{D}}$, and we can therefore write Eq. (9) in terms of the soft data (for which we write $f_{\mathrm{s}}\left(\mathbf{z}_{\mathrm{D}}\right)$ as shorthand):

$$\pi\left(z_0|f_{\mathrm{s}}\left(\mathbf{z}_{\mathrm{D}}\right)\right) \propto \int f_0\left(\boldsymbol{\theta}\right) f_{\mathrm{SRF}}\left(\mathbf{z}_{0\mathrm{D}}|\boldsymbol{\theta}\right) f_{\mathrm{s}}\left(\mathbf{z}_{\mathrm{D}}\right) \mathrm{d}\mathbf{z}_{\mathrm{D}}\, \mathrm{d}\boldsymbol{\theta}\,. \tag{11}$$

By writing the prediction distribution in this way, we may note the similarities between this Bayesian hierarchical approach and the way in which the Bayesian maximum entropy (BME; Christakos, 2000) method incorporates soft data in a spatial analysis; the predictive distribution for both methods is obtained by integrating the product of the soft data and the SRF — or in the BME terminology, the general knowledge based — model. Orton and Lark (2009) showed how the BME approach could be used to give predictions for lognormal variables using soft data; this work, and indeed the general BME methodology, is based on knowledge of the covariance and mean trend parameters. In the hierarchical approach, however, we also integrate over the trend, covariance and transformation parameters, to incorporate the uncertainty about these values.

2. Priors – The task of specifying appropriate prior distributions for the parameters is an issue of considerable interest in Bayesian statistics. The *subjective* Bayesian approach (e.g. Banerjee et al., 2004) consists of choosing prior distributions to represent our *a priori* belief about the values of the system parameters (perhaps based on the opinions of experts, or on the results of previous experiments). The *objective* Bayesian approach (e.g. Berger et al., 2001) is to determine appropriate prior distributions to use for these parameters in circumstances where such prior information is unavailable.

In this work, we use a combination of the subjective and objective approaches for our parameter priors. We make the assumption that the parameters are *a priori* independent. For the (constant) mean and total variance parameters, we adopt the commonly used improper uninformative prior (Jeffrey's independence prior; Jeffreys, 1961):

$$f_0\left(\mu,\sigma^2\right) \propto \frac{1}{\sigma^2}. \tag{12}$$

Since this prior is improper (i.e. its integral is infinite), we must ensure that it gives rise to proper posterior distributions. Note that this prior for σ^2 is equivalent to a uniform prior for $\ln\sigma^2$. Berger et al. (2001) showed that the improper prior, Eq. (12), for the mean and variance parameters gives rise to proper posterior distributions in a spatial analysis if the priors for the spatial correlation parameters, *s* and *a*, are proper. We follow De Oliveira (2005) by subjectively assigning vague proper priors for these parameters (i.e. priors that represent very little knowledge about the parameters), using an inverse Gamma distribution for the effective range parameter, *a*, and a uniform prior for the proportion, *s*. With parameters of $\alpha = 2$ and $\beta = \hat{a}$, the prior distribution for *a* has a mean of $\hat{a}$ and an infinite variance, where we choose $\hat{a} = 15$ km to represent our *a priori* guess of the range (evaluated through inspection of the experimental variogram). This gives:

$$f_0(\boldsymbol{\theta}) = f_0\left(\mu,\sigma^2,a,s\right) \propto \frac{1}{\sigma^2 a^{\alpha+1}} \exp\left\{\frac{-\beta}{a}\right\}. \tag{12}$$

We note here that the posterior distribution for the range, *a*, is influenced by the choice of prior distribution here (see results in Section 5.1). De Oliveira (2005) found similar results, and suggested that this was because the likelihood for *a* is quite flat. However, in terms of the resulting predictions, the prior does not have much

391 influence, because we have sufficient data in this case study that the posterior
392 distribution for Z_0 is dominated by the likelihood.
393
394 *3. Approximating via MCMC* – The multivariate integral in Eq. (11) is not analytically
395 tractable. However, some of the components of the parameter vector, $\boldsymbol{\theta}$, may be
396 integrated out analytically, leading to a simplification of any numerical technique that
397 we use to approximate the predictive distribution; we can integrate Eq. (11) with
398 respect to the parameters, μ and σ^2, to yield:

399 $$\pi\left(z_0 \middle| f_{\mathrm{s}}\left(\mathbf{z}_{\mathrm{D}}\right)\right) \propto \int f_{\mathrm{s}}\left(\mathbf{z}_{\mathrm{D}}\right) f_0(s,a) f_{\mathrm{ISRF}}\left(\mathbf{z}_{\mathrm{D}} \middle| s,a\right) f_{\mathrm{ISRF}}\left(z_0 \middle| \mathbf{z}_{\mathrm{D}}, s, a\right) \mathrm{d}\mathbf{z}_{\mathrm{D}}\, \mathrm{d}s\, \mathrm{d}a\,, \qquad (14)$$

400 where we refer to $f_{\mathrm{ISRF}}\left(\mathbf{z}_{\mathrm{D}} \middle| s,a\right)$ as the integrated SRF model, and $f_{\mathrm{ISRF}}\left(z_0 \middle| \mathbf{z}_{\mathrm{D}}, s, a\right)$ as
401 the related predictive distribution. We noted in Eq. (7) the relationship between the
402 distribution for **z** and that for **y**; the pdf for **z** is defined by the product of the pdf for **y**
403 and the Jacobian of the transformation, J_φ. If we assume that $\mathbf{y}_{0\mathrm{D}}$ has a multivariate
404 Gaussian distribution, then we have:

405 $$f_{\mathrm{ISRF}}\left(\mathbf{y}_{\mathrm{D}} \middle| s,a\right) \propto \frac{1}{\left|\mathbf{A}_{\mathrm{D}}\right|^{\frac{1}{2}} \left(\mathbf{1}^{\mathrm{T}} \mathbf{A}_{\mathrm{D}}^{-1} \mathbf{1}\right)^{\frac{1}{2}} \left[\mathbf{y}_{\mathrm{D}}^{\mathrm{T}} \mathbf{T}_{\mathbf{A}} \mathbf{y}_{\mathrm{D}}\right]^{\frac{N-1}{2}}}\,, \qquad (15)$$

406 and:

407 $$f_{\mathrm{ISRF}}\left(y_0 \middle| \mathbf{y}_{\mathrm{D}}, s, a\right) = \frac{\Gamma\left(\frac{N}{2}\right)}{\sqrt{\pi}\,\Gamma\left(\frac{N-1}{2}\right) \left\{N \operatorname{var} \hat{Y}_{\mathrm{OK}}\right\}^{\frac{1}{2}} \left[\frac{\left(y_0 - \hat{Y}_{\mathrm{OK}}\right)^2}{N \operatorname{var} \hat{Y}_{\mathrm{OK}}} + 1\right]^{\frac{N}{2}}}\,, \qquad (16)$$

408 where $\mathbf{A}_{\mathrm{D}}$ is the correlation matrix for the data locations, and
409 $\mathbf{T}_{\mathbf{A}} = \mathbf{A}_{\mathrm{D}}^{-1} - \frac{\mathbf{A}_{\mathrm{D}}^{-1} \mathbf{1}\mathbf{1}^{\mathrm{T}} \mathbf{A}_{\mathrm{D}}^{-1}}{\mathbf{1}^{\mathrm{T}} \mathbf{A}_{\mathrm{D}}^{-1} \mathbf{1}}$. Here we write $\hat{Y}_{\mathrm{OK}}$ for the ordinary kriging estimate and
410 $\operatorname{var} \hat{Y}_{\mathrm{OK}}$ for the ordinary kriging variance for Y_0 (where the total variance parameter,

411 or sill, σ^2, has been estimated by maximum likelihood, given the values of *s* and *a*).
412 The predictive distribution, Eq. (16), is equivalent to a *t*-distribution with $N-1$
413 degrees of freedom for the standardized variable, $\frac{y_0 - \hat{Y}_{\mathrm{OK}}}{\sqrt{N \operatorname{var} \hat{Y}_{\mathrm{OK}} \big/ N-1}}$. This predictive
414 distribution has a mean of $\hat{Y}_{\mathrm{OK}}$, and a variance of:

415 $$\sigma^2_{\mathrm{ISRF}} = \operatorname{var} \hat{Y}_{\mathrm{OK}} \frac{N}{(N-3)}. \qquad (17)$$

416 We note here that although the mean and variance of this distribution do exist for the
417 log-transformed (i.e. Gaussian) variable, *Y*, the back-transformed pdf for the original
418 variable, *Z*, does not have a defined mean or variance. It is, however, a proper
419 probability distribution, and all quantiles of this pdf can be calculated.
420 Eq. (14) thus gives the prediction distribution in a form that we may make use
421 of in its numerical approximation. This integral may be approximated by the
422 summation:

423 $$\pi\left(z_0 \middle| f_{\mathrm{s}}(\mathbf{z}_{\mathrm{D}})\right) \approx B^{-1} \sum_{i=1}^{M} f_{\mathrm{ISRF}}\left(z_0 \middle| [\mathbf{z}_{\mathrm{D}}, s, a]^{(i)}\right), \qquad (18)$$

424 where B^{-1} is a normalization constant, and $[\mathbf{z}_{\mathrm{D}}, s, a]^{(i)}$ is the *i*th of *M* independent
425 samples drawn from the probability distribution described by:

426 $$f(\mathbf{z}_{\mathrm{D}}, s, a) \propto f_{\mathrm{s}}(\mathbf{z}_{\mathrm{D}}) f_0(s, a) f_{\mathrm{ISRF}}(\mathbf{z}_{\mathrm{D}} | s, a). \qquad (19)$$

427 We may draw samples from this probability distribution using a Markov chain Monte
428 Carlo (MCMC) method, in particular the Metropolis-Hastings algorithm — see Gilks
429 et al. (1996) for a good introduction to the general theory and some applications of
430 MCMC. In any MCMC algorithm, we begin with a set of samples (one sample for
431 each of the variables in the system). A new set of samples is then drawn, which is
432 conditional on the previous set, and that set only; hence the name Markov chain

Monte Carlo. There are several classes of algorithm for performing MCMC, based on different methods of drawing the samples. In Metropolis-Hastings, new values for the set of variables (or alternatively for an individual variable) are proposed, and this set of values (or value) is accepted with a probability that depends on the joint probability, Eq. (19). For instance, the probability of accepting a proposed sample, **Y**, when the previous sample is **X**, is given by:

$$p_{\text{acc}} = \min\left(1, \frac{f(\mathbf{Y})q(\mathbf{X}|\mathbf{Y})}{f(\mathbf{X})q(\mathbf{Y}|\mathbf{X})}\right) \tag{20}$$

where $f(\mathbf{Y})$ is the joint probability density, Eq. (19), for the proposed sample **Y** and the current state of all other system variables, and $q(\mathbf{Y}|\mathbf{X})$ is the probability density for the proposed sample, **Y**, from the proposal distribution. The proposal distributions can have any form, but they should be chosen so that the resulting samples explore the posterior distribution effectively. The acceptance probabilities ensure that if a sample, $[\mathbf{z}_{\text{D}}, s, a]^{(i)}$, is a sample from the posterior distribution, then $[\mathbf{z}_{\text{D}}, s, a]^{(i+1)}$ is also a sample from this distribution. If we run the algorithm for long enough, then the chain will 'forget' its initial state, and thereafter, the samples may be considered as (dependent) samples from the posterior distribution. Consecutive samples may be highly correlated, and to reduce this correlation, we can save the samples from every *k*th iteration only.

We used five independent chains, started from five different sets of initial values. After tuning (for which we followed De Oliveira (2005) by tuning the proposal distributions to produce an acceptance rate of around 0.35) and burning in (forgetting the initial state) these chains, we saved every fifth sample from each chain, saving a total of 5000 samples from each chain. We compared the estimates and

predictions from each chain, and concluded that we could be confident that our results were accurate to the levels given in this paper.

We used MATLAB (2004) to perform the calculations. For practitioners that are interested, we can provide the MATLAB code at request. It is also possible to use the freely available WinBUGS software package (Spiegelhalter et al., 2005) to draw samples from the posterior distributions in a geostatistical case study. However, we found this to be considerably slower than MATLAB in this example.

Other issues

Some statistic of the posterior pdf for $Z(\mathbf{x}_0)$, Eq. (18), may be used as the prediction. Often, the mean is chosen so as to minimize the mean squared error (MSE). However, we cannot use the mean in our case, since — as we noted earlier — the mean of the predictive distribution, $f_{\mathrm{ISRF}}(z_0|\mathbf{z}_{\mathrm{D}}, s, a)$, does not exist. In this study, we therefore consider the median of the posterior distribution as our predictor. For heavily skewed distributions, the mean is sensitive to the variance of the transformed variable; the median predictor has often been preferred for this reason. Tolosana-Delgado and Pawlowsky-Glahn (2007) justify the use of the median predictor for lognormal variables through its property as the optimal predictor on the multiplicative scale, the scale on which the lognormal distribution is built. Some measure of the uncertainty about $Z(\mathbf{x}_0)$ can be calculated from the posterior distribution; we will use the standard 90 % confidence interval, since the variance does not provide a good measure of uncertainty for heavily skewed distributions.

Ordinary lognormal kriging deals with non-stationarity in the mean by considering that this mean is constant within a local neighbourhood of the prediction location only. In this work, we deal with non-stationarity in the total variance

parameter (through Bayesian prediction) in the same manner; if we use MCMC to draw samples from the posterior distribution for $[\mathbf{z}_D, s, a]$ given the global data, then we can use Eq. (16) within the local neighbourhood to give the conditional prediction distribution (we use Eq. (18) to compute the average of such probability densities to integrate over $\mathbf{z}_D$, *s* and *a*). By doing this, we address non-stationarity in the total variance, σ^2.

We consider Bayesian hierarchical modelling using three different approaches to represent the measurement error:

1). For comparison with OLK, we considered the detection limit approach to represent the measurement error, whereby the small measurements were set to a value of half of the DL. We used just the measurements greater than the DL to sample the covariance model parameters. (LBH)

2). We represented the small measurements by interval type soft (censored) data on $(0, DL)$. Here, we used these censored data (as well as the larger measurements) to sample the covariance model parameters, since treating them as soft data allows them to vary and hence not contribute unduly to the nugget effect. (LBC)

3). We represented each measurement by a soft datum through Eq. (3). (LBS)

5. Results

We focussed here on one particular part of the study area that roughly corresponds to the Fens region of East Anglia (the area outlined in the north-west of the region in Fig. 2); we did so for two reasons. Firstly, this region was of particular interest to us, since it contained more of the lower measurements of Se, where our treatments of the measurement error were most different. Secondly, as can be seen

from the plot of the data in Fig. 2, the data over the entire study region cannot be assumed to provide a realization of a stationary random function, a requirement of many standard geostatistical techniques; the measurements in the Fens region show a high degree of spatial smoothness (the variogram that is fitted to these data only have a proportion of the total variance with a spatial structure of $s = 0.59$), whilst the measurements over the remainder of the study area showed less spatial correlation (this variogram had a spatial variation parameter, $s = 0.15$). This difference may be explained by the history of the low-lying Fens region, which was liable to flooding — in some cases, permanently flooded — before being drained. This drainage was essentially started in the 17th century to provide farmland, although once drained, the peat that covered much of the region shrank leaving the land lower than the surrounding rivers, and by the end of the 17th century the land was once again under water (Godwin, 1978). Drainage was again attempted in the late 18th and early 19th century, and completed when the dawn of the steam age in the 1820s provided more powerful pumps to replace the windmills. The concentration of Se in the topsoil is strongly related to the quantity of soil organic carbon (SOC); the smoothness of Se concentration in the Fens and high variability in the remainder of the region is due to the different spatial distributions of SOC in these two parts of East Anglia.

For the purposes of validation, we split the Fens dataset into two parts, one for estimation and one for validation. The estimation dataset contained the measurements at 564 locations across the region, whilst the validation dataset consisted of the measurements at the remaining 1127 locations. We also used these estimation data to produce maps using the different spatial prediction methods.

5.1 Spatial correlation models

For ordinary lognormal kriging, we must fit a variogram to the log-transformed data. In our OLK approach, we use a value of half of the detection limit to represent the measurements that were not statistically different from zero. We therefore only used the measurements that were greater than the DL for fitting the variogram, since the repetition of the value, $DL/2$, contributes nothing to many of the squared differences and reduces the nugget of the variogram. The experimental and fitted model variogram are plotted in Fig. 4; the parameters for the fitted exponential model were $\sigma^2 = 0.44$, $s = 0.59$ and $a = 19\,\text{km}$.

The Bayesian approach does not require that a single variogram be fitted to the data. Instead, the method integrates over all possible variogram parameters using the integrated prediction formula, Eq. (18). We can use the samples from the posterior distributions for the correlation parameters to calculate the mean, and 5 and 95 percentiles of our estimated correlation at various lag distances. In Fig. 5, we plot these statistics for the associated normalized variogram (i.e. the variogram normalized to unit variance, since the variance parameter is integrated analytically); the three plots show the posterior statistics for the normalized variograms for the LBH, LBC and LBS approaches. Table 2 gives the posterior statistics for the variogram parameters. From these, we can see that the soft data approach gave a larger value for the spatial correlation parameter, *s*.

The results shown in Table 2 were obtained using the inverse gamma prior for the effective range parameter, *a*, (with the prior guess of $\hat{a} = 15$ km) and uniform priors for μ, s and $\ln\sigma^2$. We tested the sensitivity of the results (with the LBS approach) to the parameters of this inverse gamma prior: with a prior guess of $\hat{a} = 30$ km, the posterior statistics for *s* were unchanged, whilst the posterior mean for *a* was increased to 30 km, and the 90 % CI became $[17,52]$. These differences did not affect

556 the spatial predictions or the estimates of uncertainty that we consider in the following
557 sections. This was because in this case study we have enough data so that they
558 dominate the posterior through the likelihood; if we had fewer data, then the prior
559 distribution may have been of more importance. We tested the sensitivity of the
560 results to the prior guess for the effective range parameter with 100 estimation data.
561 With the LBS approach, and with prior guesses of $\hat{a} = 8$ km, $\hat{a} = 15$ km and $\hat{a} = 30$
562 km, the posterior means for *a* were 37 km, 39 km and 44 km, respectively. The
563 resulting LBS predictions were similar, with a maximum absolute difference between
564 predictions from the three different priors of 0.05, and identical (to three significant
565 figures) validation results for the bias, MSE and GMSE from all three priors.
566
567 *5.2 Maps of geostatistical predictions*
568 We used the methods to estimate the Se concentration at the nodes of a grid
569 that covered the area of interest in our case study. For OLK, we compared the maps
570 produced using the mean and the median predictors, whilst for the remaining maps,
571 we used the median predictor only. We found that for the Bayesian methods, 1000
572 samples from the posterior distributions for the model parameters were enough to give
573 sufficient accuracy for these maps (i.e. there were no visual differences between maps
574 produced using 1000 samples). The maps are shown in Fig. 6.
575 Fig. 6a shows maps for the OK predictor, and for the mean and median OLK
576 predictors. The OK map shows the largest area of dark (i.e. the high predicted Se
577 concentrations). This is because the skew of the data is not taken into account by OK,
578 and therefore the large data values have a strong influence on the predictions. When
579 the data are transformed (i.e. by taking logarithms) the largest measurements do not
580 have such a great influence on the predictions. By comparing the maps for the mean

581 and median OLK predictors, we can see that the mean gives larger predictions, as
582 should be expected.
583 If we look at the maps produced using the lognormal Bayesian method (i.e.
584 LBH, LBC and LBS, Fig. 6b), then we can compare our different treatments of
585 measurement error. The features of these maps all appear very similar; the dark (i.e.
586 the soil with a predicted Se concentration of more than 0.5 mg kg $^{-1}$) in each of the
587 maps covers roughly the same parts of the region. However, we can notice small
588 differences. The map produced using the soft data shows generally slightly larger
589 predictions (compare the sizes of the dark regions, and also the areas of lighter shades
590 of grey), and is also smoother than the maps produced using the censored or hard
591 data. This is because of the effect that the hard (and censored) data have on the
592 predictions. When we impute a value of half of the DL for the smaller measurements
593 in the LBH approach, this datum is allowed to have a larger influence on the
594 predictions than it should really have, since the uncertainty about this imputed value is
595 not accounted for. This causes the predictions around the imputed values to be
596 smaller. Similarly, the censored data approach gives lower predictions than the soft
597 data approach; the representation of a measurement of 0.2 mg kg $^{-1}$ by the interval
598 [0,0.2] does not allow for the measurement error to give a true concentration of Se
599 greater than 0.2 mg kg $^{-1}$. The soft data approach, on the other hand, aims to represent
600 exactly the information that each single measurement provides (about the
601 concentration at that measurement location) through the soft data pdfs. For example,
602 even though measurements of 0.2 mg kg $^{-1}$ and 0 mg kg $^{-1}$ are not statistically different
603 from each other, the soft data pdfs can be used to represent these measurements
604 differently, by taking into account the uncertainty about each different measurement.
605 In our opinion, the smoother transition between the larger and smaller predictions —

606 as modelled by the soft data approach — provides a better representation of the
607 uncertainty in the predictions surrounding the lower measurements.
608 We can investigate the effects of parameter uncertainty on the predictions by
609 comparing the map produced using OLK (the median predictor) to that produced
610 using the lognormal Bayesian method with the same hard data (i.e. LBH). These two
611 maps, shown in Fig. 6c, appear very similar. Both approaches are based on the same
612 model for the SRF; the only difference between these approaches is that the Bayesian
613 approach accounts for the uncertainty about the variogram parameters, and thus it
614 would appear that this parameter uncertainty is not of great importance for the
615 predictions in this case study.
616
617 *5.3 Prediction assessment*
618 We have compared the methods in terms of the maps of the resulting
619 geostatistical predictions. We can also use validation to compare the predictions. We
620 estimated the Se concentration at the 1127 validation locations using the geostatistical
621 approaches described in this paper. We then compared the predicted values to the
622 actual measurements at these sites in terms of the bias, mean squared error (MSE),
623 and geometric MSE (GMSE). We used both the arithmetic and geometric means of
624 the squared errors because the MSE for lognormal predictions is dominated by errors
625 at just a few locations (i.e. when we under-predict the concentration by a large
626 amount), whilst the GMSE is effectively a measure of the errors on the log-
627 transformed scale; since our data are roughly Gaussian on the logarithmic scale, the
628 GMSE is not dominated by these errors.
629 We note here that when we use the Bayesian method based on the pdf-type
630 soft data (i.e. LBS), the variable that we are actually predicting is $Z(\mathbf{x}_0)$. However,

631 the values that we are validating against are measurements of this variable. We can
632 put the measurement error back into the predictions from LBS through:

633 $$\pi\left(\zeta_0 \middle| f_s\left(\mathbf{z}_D\right)\right) = \int \pi\left(z_0 \middle| f_s\left(\mathbf{z}_D\right)\right) f_{me}\left(\zeta_0 \middle| z_0\right) dz_0 . \quad (21)$$

634 We therefore use this prediction pdf to calculate the predictions for validation with the
635 LBS approach. Since the approaches built on the detection limit (i.e. the hard and
636 censored data approaches) do not explicitly distinguish between the measured and
637 actual values of the variable, we could not consider such an approach with these
638 methods. We therefore use our original predictions with these methods.

639 Table 3 summarizes the results for the estimators in terms of the bias, MSE,
640 and GMSE. This shows the results for the validation against the actual measurements,
641 and the GMSE for validation against the DL-imputed values; the results for validation
642 against the soft data means showed similar patterns to the validation against the actual
643 measurements, and are not shown.

644 The OK predictor gave the best predictions of all of the methods in terms of
645 the MSE and bias of the predictions. However, as noted previously, the MSE is
646 dominated by the errors at the few locations where the measured concentration is in
647 the tail of the distribution. At these locations, there is no danger of us over-predicting
648 and so the larger the prediction, the better. Generally, OK performs well (in terms of
649 the MSE and bias) because when we krige with the original (untransformed) data, the
650 larger measurements have a big effect on the predictions (since in this model they are
651 essentially outliers). When we use the log-transformed data to predict, the larger
652 measurements do not act as outliers, and do not affect the prediction as much,
653 resulting in a lower prediction. Hence, we get a larger error at these locations — and
654 therefore also a larger MSE — from a method built on this transformation. Because of
655 this domination of the MSE (and bias) by the errors at just a few locations, the GMSE

656 provides a more appropriate measure of the accuracy of the geostatistical predictions
657 in this case study. We therefore focus on this measure from here on. We can see that
658 despite giving the best predictions in terms of the MSE, OK gave the worst in terms
659 of the GMSE.

660 Tolosana-Delgado and Pawlowsky-Glahn (2007) justify the use of the median
661 predictor (for lognormal data) based on its property as the optimal predictor on the
662 multiplicative scale; the GMSE is essentially a measure of the errors on this scale. As
663 should be expected, we see that the median OLK predictor gives a smaller GMSE (but
664 larger MSE) than the mean OLK predictor (whatever the choice of validation values).
665 The OLK mean predictor has often been disregarded because of its sensitivity to the
666 fitted variogram parameters. The predictor aims to achieve unbiased predictions (and
667 minimize the MSE on the logarithmic scale) through a balance between many small
668 over-predictions and a few large under-predictions; this balance is sensitive to the
669 fitted variogram parameters and to the lognormal assumption, from which any
670 departure can result in poor predictions (Roth, 1998). The median predictor
671 overcomes this sensitivity somewhat, because the back-transform does not depend
672 directly on the fitted variogram parameters.

673 The validation results for the Bayesian approach suggest that the pdf-type soft
674 data improves the accuracy and precision of the estimates. The LBS approach gave a
675 GMSE of 0.012, compared to 0.014 for the LBH and LBS approaches. It also gave a
676 lower bias than these other two approaches. The Bayesian approach offers the
677 advantage over the OLK approach of incorporating parameter uncertainty in the
678 predictions. The validation results, however, did not show significant differences
679 between these approaches in terms of the bias, MSE and GMSE (compare the results
680 from the median OLK predictor with those from LBH).

681 We note that the choice of prior did not affect the accuracy of the predictions;
682 we compared the results from LBS with the inverse Gamma prior for *a* and with prior
683 guesses of $\hat{a} = 15$ km and $\hat{a} = 50$ km, and found all of the measures of prediction
684 performance to be identical to the precision shown here.
685
686 *5.4 Uncertainty assessment*
687 We also used the validation dataset to determine how well each of the methods
688 represented the uncertainty about the estimated Se concentration. Consider a
689 validation location, $\mathbf{x}_{\mathrm{v}}$, where we have a measurement, ζ_{v}, and suppose that we have
690 used our estimation dataset to calculate a prediction pdf for this measurement, from
691 which we can calculate any quantile. If the prediction pdf provides a good
692 representation of the uncertainty, then we should expect that the proportion of
693 locations for which the validation measurement is less than the *q*-quantile from the
694 prediction pdf be *q*; we denote this actual proportion p_q. If we assume that the
695 validation sites are independent for $n_{\mathrm{k}} = 1127$ validation locations, then the 90 % CI
696 for p_q is $q \pm 1.645\sqrt{q(1-q)/n_{\mathrm{k}}}$. We note that the assumption of independent
697 validation data results in a somewhat crude estimate of the confidence intervals for
698 p_q. In reality, these intervals should be wider, because of the correlation between the
699 data at the validation locations. We use the independence assumption here to provide
700 a rough idea of plausible values for p_q, rather than to accept or reject a particular
701 approach based on these bounds. We display the results as plots of the quantile of the
702 prediction distribution, *q*, on the *x*-axis, against the proportion of validation
703 measurements less than this quantile, p_q, on the *y*-axis. These plots are shown in Fig.
704 7.

We also calculate the percentage of the validation locations for which the 90 % confidence interval contained the validation data for each prediction method, P_{CI}, and the average widths of these confidence intervals, W_{CI}. These are shown in Table 4. A method that gives a small average width is precise, whilst one that gives a percentage in the confidence interval close to 90 % is accurate (in terms of the uncertainty estimate).

Again, when we use the pdf-type soft data to calculate the prediction distributions (i.e. LBS), we can put the measurement error back into the prediction pdfs through Eq. (21) and use this pdf for validation; Table 4 shows the results from these pdfs, and in brackets the mean width of the 90 % CIs for the actual concentration, without the measurement error added back. We could not consider such an approach to separate out the measurement error and micro-scale variation with the hard data or censored data approaches.

From Fig. 7a, we can see that OK did not represent the uncertainty well. The lower quantiles of the prediction distribution were too low, and therefore very few validation measurements fell below these lower quantiles. The upper quantiles were too high, and very few validation measurements were greater than these upper quantiles. Also, from Table 4, we can see that the 90 % CI captured the validation measurements too often. This is because the OK estimate is symmetric (i.e. based on a Gaussian SRF) and does not take account of the highly skewed nature of the data. Note also the average width of the CIs from OK, which was much larger than that from any other method, showing that the approach overestimated the uncertainty of the predictions.

Ordinary lognormal kriging (OLK) does account for the skew of the data, but does not account for any uncertainty in the variogram. We therefore expected it to

improve on the OK estimates, but to underestimate the uncertainty of the predictions. We expected that this underestimation of uncertainty would be a result of the lower quantiles being too high, and the upper quantiles being too low. Indeed, from Fig. 7a, we can see that the upper quantiles of the OLK prediction distribution were generally too low. However, the lowest quantiles ($q \leq 0.1$) gave good representations of the uncertainty. This happened because the value imputed for the lower measurements by the DL approach (i.e. 0.1) generally underestimated the actual measurement at these locations (the mean of the actual estimation measurements at these locations was 0.15). When OLK is used to calculate the prediction distribution with these data, the variogram uncertainty is not taken into account (meaning that the lower quantiles are higher than they should be). The overestimation of the lower quantiles by OLK and underestimation of the lower quantiles by the DL approach balances out, and results in good estimates for these lower quantiles by OLK with the DL approach in this case study.

When we look at the 90 % CI for OLK, we see that this failed to capture the validation value in enough cases ($P_{\mathrm{CI}} = 80.8$ %). Generally, this was because the estimated CIs were too narrow; the average width of the CIs from OLK was the smallest out of all of the approaches. This was because OLK does not account for the uncertainty in the estimated variogram, which can play a significant part in the uncertainty of lognormal predictions.

The Bayesian approach incorporates variogram uncertainty. When we used the DL approach to give hard data (i.e. LBH), this gave better results for the upper quantiles (where the effect of the DL imputed data was less) than OLK, but worse results for the lower quantiles. This was because the underestimation of the lower quantiles by the DL approach was not balanced out by an overestimation from the

755 geostatistical approach here. The uncertainty of the variogram is accounted for by the
756 Bayesian approach, and thus, because the DL imputed data underestimate the
757 measurements, this results in lower quantiles that are lower than should be expected.
758 We note that the results from the Bayesian approach with censored data, LBC, were
759 very similar to those from LBH.

760 In terms of the 90 % CI, we can see that LBH performed better than OLK,
761 with 90.1 % of the validation measurements contained in the intervals. These intervals
762 were larger than those from OLK, because they take into account the uncertainty
763 about the variogram.

764 The Bayesian approach gave better results when we used soft data. We can see
765 that the line for LBS on Fig. 7b lies closer to the diagonal, $p_q = q$, than the line for
766 LBH. This is because the soft data better represent the information that we receive
767 from the measurements than the DL approach does through hard data. However, we
768 again see that the Bayesian approach resulted in the lower quantiles of the prediction
769 distributions being too low. Although the soft data improves on the hard data
770 approach, it could perhaps be improved further by considering an alternative
771 measurement error model; this would generally be a more complicated model, which
772 we would only be able to consider if we had more repeated measurements.

773 When we look at the 90 % CIs, the soft data approach, LBS, resulted in the
774 validation measurements being captured in these intervals more often than should be
775 expected. This was because only 2.5 % of the validation measurements fell below the
776 0.05-quantiles of the prediction distributions (94.9 % of the validation measurements
777 fell below the 0.95-quantiles). With LBH, although 90.1 % of the measurements were
778 in the 90 % CI, this was made up of 3.5 % below the 0.05-quantiles, and 93.7 %
779 below the 0.95-quantiles; since both of these are less than should be expected, the

780 resulting 90 % CIs contain the validation measurements for an acceptable number of
781 validation locations.
782 An interesting point here is that the average width of the CIs from LBS was
783 smaller than those from LBH and LBC. This was because the hard or censored data
784 caused the prediction pdfs to favour lower values than the soft data approach (because
785 the average of the replaced measurements in the LBH and LBC approaches was
786 greater than 0.1); the lognormal assumption then leads to narrower CIs from the LBS
787 approach. Further, the average width of the CIs from LBS without the measurement
788 error added back in (i.e. predictions for the actual underlying concentration, and not
789 the measurement of this quantity, shown in brackets in Table 4) is considerably
790 smaller again. This provides a good benefit of using the soft data approach — we can
791 separate out the micro-scale and measurement error components of the variation, and
792 use this information to reduce the uncertainty about our predictions.
793 We note that the choice of prior again did not affect the predictions; we
794 compared the results from LBS with the inverse Gamma prior for a and with prior
795 guesses of $\hat{a} = 15$ km and $\hat{a} = 50$ km, and found the plot of p_q against q to be
796 identical.
797
798 *5.5 Geostatistics for the effective management of selenium deficient soils*
799 The management of the soil can be made more efficient by using the
800 information provided by geostatistical predictions. In many case studies concerning
801 the concentration of some element in the soil, the task is to determine areas where the
802 soil may be considered as contaminated, and some clean-up operation may be deemed
803 necessary in these areas. In this case study, however, we concern ourselves with the
804 problem of determining areas where the soil may be considered as Se deficient; Se

805 may be added to the soil in these areas to increase the amount available for uptake by
806 plant roots.

807 We note that the variable in this case study, the total Se in the soil, is a poor
808 indicator of the total Se available to plants. Other factors, such as the Se speciation in
809 soil, the soil pH, and the sulphate concentration can have a much greater influence on
810 Se uptake. However, a limit of 0.5 mg kg $^{-1}$ is used in New Zealand, below which the
811 Se content of the grass may be insufficient for grazing sheep (Hawkesford and Zhao,
812 2007). Tan (1989) defines the level of Se in soil for human nutrition as being deficient
813 for less than 0.125 mg kg $^{-1}$, and marginal for 0.125–0.175 mg kg $^{-1}$. In this work, we
814 consider three limits (which we refer to as the limit of deficiency, or z_{D}); the first of
815 0.55 mg kg $^{-1}$, a second of 0.35 mg kg $^{-1}$, and a third of 0.15 mg kg $^{-1}$, so these limits
816 were chosen to demonstrate the differences between the geostatistical approaches in
817 this case study.

818 In order to decide whether a site has sufficient Se, we should take into account
819 the relative cost of wrongly declaring a site as Se deficient, ω_1, the relative cost of
820 wrongly declaring a site as not deficient, ω_2, and the (estimated) probabilities of these
821 events occurring. If we used the perfect strategy (i.e. correctly classified the soil at
822 each location in the validation dataset), then we would incur some minimum cost; we
823 suppose (without loss of generality) that this minimum cost is zero. When we base our
824 strategy on the probabilities of deficiency, p_{D}, estimated using our geostatistical
825 method, we incur a greater cost than this minimum due to misclassification. At any
826 single location, the expected extra cost if we apply Se is $\omega_1(1-p_{\mathrm{D}})$, and the expected
827 extra cost if we do not apply is $\omega_2 p_{\mathrm{D}}$. We choose our strategy at each location to give
828 the smaller expected cost.

We can use our validation data to calculate the costs that result from the decision about where to apply Se to the soil. We present the resulting costs in Fig. 8 as a percentage of a default 'maximum' cost. This default is the cost that would be incurred if we simply used the percentage of deficient data in the estimation dataset to give p_D and hence determine the appropriate strategy — this default tells us to apply everywhere if the cost ratio, $\omega_R = \omega_1/\omega_2$, is less than $\frac{p_D}{1-p_D}$, and to apply nowhere otherwise. Percentages of this default below 100 % indicate the potential saving (i.e. 90 % equates to a 10 % saving) that could be made by using the geostatistical technique to decide where to apply. The resulting validated percentage depends only on the ratio of the costs, $\omega_R = \omega_1/\omega_2$. We can rearrange the cost inequality that we use to choose our strategy to show that we are essentially using the $\frac{\omega_R}{1+\omega_R}$-quantiles from the prediction distributions to classify the soil as deficient or otherwise. We therefore plot this variable on the *x*-axis in Fig. 8. Quantiles below $q = 0.5$ are used when the cost of wrongly declaring the soil as deficient, ω_1, is small (and hence, this favours the application of Se to the soil), and the upper quantiles are used when this cost is large.

In Fig. 8a, we compare the methods for using hard data to calculate the geostatistical estimates (via the DL approach). The results here seem to agree with the results from the previous section — for the low quantiles, OLK performs well, but for the higher quantiles performs poorly compared to LBH. This was particularly the case for $z_D = 0.55$ ppm and for $z_D = 0.35$ ppm.

In Fig. 8b, we compare the treatments of the measurement error through the lognormal Bayesian methods, LBH, LBC and LBS. We saw in the previous section

852 that LBS generally gave better estimates of the uncertainty about the predictions; we
853 should therefore expect this to lead to better decisions regarding the management of
854 the soil. From Fig. 8b, we can see that the costs resulting from the three geostatistical
855 approaches were similar for many values of the cost ratio, particularly for the limits of
856 deficiency of $z_D = 0.35\,\text{ppm}$ and $z_D = 0.55\,\text{ppm}$. For these limits of deficiency, LBS
857 resulted in a cost that was less than the default strategy (i.e. percentages were less
858 than 100 %) for most quantiles from the prediction distribution (apart from the
859 extreme upper quantiles); LBH and LBC, on the other hand, both resulted in costs
860 greater than this default strategy for some of the upper quantiles (i.e. by using these
861 approaches to choose the appropriate strategy, we actually increase the cost over the
862 default strategy). In the previous section, we saw that the upper quantiles from the
863 LBS approach represented the uncertainty better than those from the LBH and LBC
864 approaches, and the reduction in cost from LBS when these upper quantiles are used
865 to classify the soil is a result of this improvement. For $z_D = 0.15\,\text{ppm}$, LBS generally
866 performed better than LBH and LBC, although in this case, for quantiles, *q*, between
867 0.31 and 0.63 (i.e. values of the cost ratio, ω_1/ω_2, between 0.44 and 1.7), LBS gave a
868 greater cost than the default strategy. However, this increased cost was small
869 compared to those from LBH and LBC, and these methods increased the cost over a
870 larger range of values for *q* (for LBH, this increase was for values of *q* between 0.28
871 and 0.76).
872
873 **6. Discussion and conclusions**
874 In this paper, we have compared several geostatistical approaches for
875 predicting the Se concentration in the soil using data that are subject to measurement
876 error. Environmetric datasets that consist of measurements subject to a detection limit

877 are commonplace, and we have investigated a method to represent these
878 measurements through soft data that provides more information than the approaches
879 built on hard or censored data.
880 We have focussed on the median predictor throughout this paper because it is
881 the optimal predictor (for lognormal variables) on the log-transformed (i.e.
882 multiplicative) scale. This follows other work (e.g. Tolosana-Delgado and
883 Pawlowsky-Glahn, 2007) that has suggested that this predictor be used for positive
884 variables. Further, we assessed the accuracy of the predictors in this paper by the
885 geometric mean of the squared errors for similar reasons (the arithmetic mean of the
886 squared errors is dominated by the errors at just a few locations for heavily skewed
887 positive variables).
888 We have compared the detection limit approach that has previously been used
889 (Woodside and Kocurek, 1997) with a soft data approach based on a Gaussian
890 measurement error model. Each soft datum represents the information that a
891 measurement is providing us with about the actual Se concentration at that location;
892 the uncertainty about this value is taken into account, and thus measurements that are
893 not statistically different can give rise to different soft data.
894 We found that the soft data approach generally resulted in slightly larger
895 predictions, and also smoother maps of these predictions. In our opinion, these
896 smoother predictions provide a better representation of the measurement error and the
897 resulting uncertainty. From the validation exercise, we found that the soft data
898 approach to incorporate measurement error improved the precision and accuracy of
899 the predictions compared to the classically used approach of using a detection limit.
900 When we used Bayesian modelling to calculate the prediction distributions, the soft
901 data gave a better representation of the prediction uncertainty, as shown in the p_q -

plots in Fig. 7. Although the hard and censored data approaches gave better results in terms of the proportion of validation measurements captured by the 90 % CIs, this was because of a balancing act between the overestimation of both the lower and upper quantiles by LBH and LBC. We also showed the soft data approach to generally result in better management of Se deficient soils; by taking into account the cost of Se deficient soil, we found that the LBS approach generally resulted in a lower validated total cost.

One particular benefit of the soft data approach is that it allows us to separate out the measurement error from the 'micro-scale' variance. We can use this information to calculate predictions for the underlying variable, and also (if required for validation), predictions for the measurements of this variable.

The soft data based on the classical measurement error model that we have used in this work effectively represents the simplest choice of model, given that the measurements are unbiased with a constant measurement error variance. However, if more repeated measurements were available from other samples, then it may be possible to consider a more complex model for the measurement error. For instance, it may be that when the actual concentration in a sample is low, the measurements have a low variance, whilst the measurements of higher concentrations may be more variable. With more repeated measurements, it may be possible to fit such a model and use it to give soft data; however, with the limited number of repeated measurements that we had, we could only consider the simplest choice, using the classical measurement error model.

We compared the Bayesian and 'plug-in' kriging approaches in order to assess whether parameter uncertainty had an effect on the predictions. We found that both approaches (when using the hard data only) resulted in maps that were apparently

identical. Furthermore, the validation results from the OLK median predictor were very similar to those from the LBH median predictor (i.e. the Bayesian approach using hard data), and we conclude that the parameter uncertainty did not affect the geostatistical predictions much. However, in terms of the assessment of the prediction uncertainty, the parameter uncertainty did have an effect. The 90 % confidence intervals from the Bayesian methods captured the validation measurement in close to 90 % of the validation cases, whilst the OLK CIs captured the validation values in only around 80 % of the cases. The improvement may be attributed to the sensitivity of lognormal kriging estimates to the fitted variogram parameters. The Bayesian confidence intervals were on average larger than those from OLK (1.16 for LBH compared to 0.76 for OLK), because of the effect that parameter uncertainty has on the uncertainty of geostatistical predictions. Although OLK poorly represented the 90 % CIs, it did represent the lower quantiles well. The hard data imputed for the low measurements (i.e. half of the DL) provide underestimates of the actual values here; also, OLK does not incorporate variogram uncertainty, and hence overestimates these lower quantiles of the prediction distribution. We hypothesized that these two poorly represented quantities could balance out in this case study to give good representations of the lower quantiles. We note that the upper quantiles were poorly represented by OLK.

In this paper, we have used the Bayesian hierarchical approach to deal with measurement error in lognormal variables. The measurement error model describes how the measurements are related to the SRF; we have referred to these measurements as soft data, because they are related to the SRF by a stochastic relationship. In fact, any information of this form may be viewed as soft data, and would fit in with the hierarchical approach. For instance, it might be that a process

model gives us information about the SRF at some locations; this will generally be uncertain information, and if its relationship to the underlying SRF can be modelled, then we can incorporate this through soft data in the hierarchical approach. We note that there are other approaches for including information from process models in geostatistical analyses (e.g. Stacey et al., 2006); the soft data approach provides another means of incorporating this information.

It is also important that we consider a hierarchical approach for the covariance parameters when we have soft data in order to ensure that the covariance parameters that we estimate are for the same variable that the soft data provides information about. For instance, if we used the mean of each soft datum to estimate the variogram, then the estimated parameters would not incorporate the measurement error properly; the nugget effect for the variogram of the SRF should not include the measurement error here, because this is accounted for in the soft data, and an estimate of the parameters based on the means of the soft data would therefore be an overestimate.

According to Deutsch and Journel (1992, p. 58), "… it is subjective interpretation … that makes a good model; the data by themselves, are rarely enough". This would seem to provide a good argument in favour of using soft data to model measurement error; all of the measurements can be interpreted through the classical measurement error model. This approach enables us to separate out the nugget variation into components for the micro-scale variation and the measurement error, which can improve the accuracy of predictions and provide the modeller with more useful information about the variability of the spatial random field.

Acknowledgements

This work was funded by the Biotechnology and Biological Sciences Research Council of the United Kingdom through its Core Strategic Grant to Rothamsted Research. This paper is published with the permission of the Executive Director of the British Geological Survey (Natural Environment Research). We acknowledge the contributions of all staff from the British Geological Survey involved in the soil survey of East Anglia: (i) the G-BASE project staff who organized the collection and processing of the soil survey data, (ii) the staff who prepared the samples, and (iii) the analytical staff who did the XRF-S analyses. We also thank Fang-Jie Zhao from the department of soil science at Rothamsted for his help with the section on selenium deficiency.

References

Adams, M. L., Lombi, E., Zhao, F.J., McGrath, S. P. 2002. Evidence of low selenium concentrations in UK bread-making wheat grain. Journal of the Science of Food and Agriculture 82, 1160-1165.

Banerjee, S., Carlin, B.P., Gelfand, A.E. 2004. Hierarchical Modelling and Analysis for Spatial Data. Chapman & Hall/CRC, Boca Raton, Florida.

Berger, J.O., De Oliveira, V., Sanso, B. 2001. Objective Bayesian Analysis of Spatially Correlated Data. Journal of the American Statistical Association 96, 1361–1374

1002 Christakos, G. 2000. Modern spatiotemporal geostatistics. Oxford University Press,
1003 New York.
1004
1005 De Oliveira, V. 2005. Bayesian Inference and Prediction of Gaussian Random Fields
1006 Based on Censored Data. Journal of Computational and Graphical Statistics 14,
1007 95–115
1008
1009 Deutsch, C.V., Journel, A.G. 1998. GSLIB: Geostatistical Software Library and
1010 User's Guide. Oxford University Press, New York.
1011
1012 Fan, M.-J., Zhao, F.-J., Poulton, P.R., McGrath, S.P. 2008. Historical changes in the
1013 concentrations of selenium in soil and wheat grain from the Broadbalk experiment
1014 over the last 160 years. Science of the Total Environment 389, 532–538
1015
1016 Fridley, B.L., Dixon, P. 2007. Data augmentation for a Bayesian spatial model
1017 involving censored observations. Environmetrics 18, 107–123
1018
1019 Gilks, W.R., Richardson, S., Spiegelhalter, D.J. 1996. Markov Chain Monte Carlo in
1020 Practice. Chapman & Hall/CRC, Boca Raton, Florida.
1021
1022 Godwin, H. 1978. Fenland: Its Ancient Past and Uncertain Future. Cambridge
1023 University Press, Cambridge.
1024
1025 Hawkesford, M.J., Zhao, F.-J. 2007. Strategies for increasing the selenium content of
1026 wheat. Journal of Cereal Science 46, 282–292

Jeffreys, H. 1961. Theory of Probability. Oxford University Press, London.

Journel, A.G. 1980. The Lognormal Approach to Predicting Local Distributions of Selective Mining Unit Grades. Mathematical Geology 12**,** 285–303

Kapur, J.N., Kesavan, H.K. 1992. Entropy Optimization Principles with Applications. Academic Press, London.

Lark, R.M., Bellamy, P.H, Rawlins, B.G. 2006. Spatio-temporal variability of some metal concentrations in the soil of eastern England, and implications for soil monitoring. Geoderma 133, 363–379

Laslett, G.M., McBratney, A.B. 1990. Estimation and implications of instrumental drift, random measurement error and nugget variance of soil attributes – a case study for soil pH. Journal of Soil Science 41, 451–471

Orton, T.G., Lark, R.M. 2009. The Bayesian maximum entropy method for lognormal variables. Stochastic Environmental Research and Risk Assessment 23, 319–328

Reimann, C., Filzmoser, P. 2000. Normal and lognormal data distribution in geochemistry: death of a myth. Consequences for the statistical treatment of geochemical and environmental data. Environmental Geology 39, 1001–1014

1051 Roth, C. 1998. Is Lognormal Kriging Suitable for Local Estimation? Mathematical
1052 Geology 30, 999–1009
1053
1054 Spiegelhalter, D., Thomas, A., Best, N., Lunn, D. 2005. WinBUGS User Manual,
1055 Version 1.4, MRC Biostatistics Unit, Institute of Public Health and Department of
1056 Epidemiology & Public Health, Imperial College School of Medicine. Available
1057 at http://www.mrc-bsu.cam.ac.uk/bugs
1058
1059 Stacey, K.F., Lark, R.M., Whitmore, A.P., Milne, A.E. 2006. Using a process model
1060 and regression kriging to improve predictions of nitrous oxide emissions from
1061 soil. Geoderma 135, 107–117
1062
1063 Tolosana-Delgado, R., Pawlowsky-Glahn, V. 2007. Kriging Regionalized Positive
1064 Variables Revisited: Sample Space and Scale Considerations. Mathematical
1065 Geology 39, 9529–558
1066
1067 Webster, R., Oliver, M.A. 2007. Geostatistics for Environmental Scientists: second
1068 edition. Wiley & Sons, Chichester.
1069
1070 Woodside, G., Kocurek, D. 1997. Environmental, Safety, and Health Engineering.
1071 Wiley & Sons, New York.

Figure captions

Fig. 1 – The marginal distribution of a) Se concentrations and b) the log-transformed Se concentrations from the entire East Anglia region

Fig. 2 – A map showing the measured concentrations of Se across the East Anglia region, and the outline of the Fens (the area in the north-west of the region) studied in this work

Fig. 3 – A graphical representation of the Bayesian spatial prediction approach

Fig. 4 – The experimental and model (exponential) variogram fitted to the log-transformed data using only the measurements greater than the DL

Fig. 5 – The posterior means, 5 and 95 percentiles from the standardized variogram using the three Bayesian approaches, LBH, LBC and LBS. In a) we only use the measurements greater than the DL to calculate the posterior variogram, in b) we also use the censored data representation for the low measurements, and in c) we use the soft data representation for all of the measurements

Fig. 6 – Maps of the geostatistical predictions of Se concentration across the Fens. In plot a), we compare OK, and the OLK mean and median predictors, in plot b), we compare the representations of measurement error by the LBH, LBC and LBS approaches, and in plot c) we investigate the effects of parameter uncertainty by comparing OLK and LBH

Fig. 7 – The proportion, p_q, of validation data less than the *q*-quantiles from the prediction distributions. Plot a) compares OK, OLK and LBH, and plot b) compares LBH, LBC and LBS. The dots show the expected value for p_q and the crosses the 90 % CIs for p_q

Fig. 8 – The cost of Se deficient soils calculated for the validation locations. The costs are presented here as a percentage of the 'default' cost, which is the cost that would be incurred if we did not use geostatistics to classify the soil. The variable on the abscissa is the quantile of the prediction distribution that is used to classify the soil. Plot a) compares OK, OLK and LBH, and plot b) compares LBH, LBC and LBS