Optimized multi-phase sampling for soil remediation surveys

B. P. Marchant*[a,b], A. B. McBratney[c], R. M. Lark[a,b], B. Minasny[c].

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

[b] *British Geological Survey, Keyworth, Nottingham, NG12 5GG, UK*

[c] *Faculty of Agriculture, Food and Natural Resources, The University of Sydney, NSW 2006, Australia*

* Corresponding author: B.P. Marchant

British Geological Survey

Keyworth, Nottingham, NG12 5GG

UK

phone: +44 (0)115 9363100

e-mail: benmarch@bgs.ac.uk

Highlights

H1: Algorithm to optimize sampling locations in multi-phase soil remediation surveys

H2: Minimizes expected costs of misclassifying local contamination status

H3: Number of samples are selected using forecasts of effectiveness of proposed designs

H4: Non-Gaussian properties are represented by copula-based models

H5: Later phases concentrate observations where contamination status is most uncertain

Abbreviations

AIC: Akaike information criterion; AEIL: Australian environmental investigation limit; EBLUP: Empirical best linear unbiased predictor; ML: Maximum likelihood; pdf: probability density function; SSA: Spatial simulated annealing

Abstract

We develop an algorithm to optimize the design of multi-phase soil remediation surveys. The locations of observations in later phases are selected to minimize the expected loss incurred from misclassification of the local contamination status of the soil. In contrast to existing multi-phase design methods, the location of multiple observations can be optimized simultaneously and the reduction in the expected loss can be forecast. Hence rational decisions can be made regarding the resources which should be allocated to further sampling. The geostatistical analysis uses a copula-based spatial model which can represent general types of variation including distributions which include extreme values. The algorithm is used to design a hypothetical second phase of a survey of soil lead in Glebe, Sydney. Observations in this phase are generally dispersed on the boundaries between areas which according to the first phase either require, or do not require, remediation. The algorithm is initially used to make remediation decisions at the point scale but we demonstrate how it can be used to inform over blocks.

1. Introduction

Human-health and environmental concerns require the remediation of contaminated soils

near former industrial sites throughout the world. In many cases, thresholds have been defined for the permissible concentration of metals and other contaminants in the soil (e.g. [1]). If the contamination is localized then spatial surveys can be conducted to suggest where concentrations are greater than these thresholds and hence remediation is required [2]. Uncertainty is inevitably attached to the results of such surveys and geostatistical techniques are used to assess the probability that a particular location is falsely designated as contaminated or not contaminated. This information, combined with an understanding of the costs of an incorrect remediation decision at a site, permit an informed decision about the extent of the remediation.

The accuracy and cost of soil-remediation surveys increase with the number of observations made. It has previously been suggested (e.g. [2], [3], [4], [5]) that the efficiency of surveys can be improved if they are split into a number of phases. The initial phase yields a low-resolution map of soil contamination. It might show that no further measurements are required in much of the study area where the soil can be designated with great certainty as either contaminated or not contaminated. The later phases concentrate observations in parts of the study region where the contamination status is in doubt. As the survey progresses the resolution of the contamination map in these regions increases until eventually it is suitable to select the locations which are to be remediated. Heuvelink et al. [6] consider a related problem in the design of mobile radioactivity monitoring networks. Normally the network is fairly coarse but in the event of a nuclear accident more sensors are required close to the accident site.

There are two key issues to address before such a multi-phase strategy can be used in practice. The first is the amount of additional sampling. How many observations should

be made, how should they be divided between phases and how should the practitioner decide when a survey is adequate? The second issue is the selection of the locations of observations within a single phase of the survey. We consider the situation where a phase of sampling has been conducted and kriging [7] has been used to predict the contamination across the study region. Two factors dictate whether further sampling is advantageous at a particular location $\mathbf{x}$. The first is how close the local prediction of the soil contamination $\widehat{z}(\mathbf{x})$ is to the threshold z_c. The second is the uncertainty of this prediction. This uncertainty can be expressed in terms of the kriging variance $\sigma^2(\mathbf{x})$. Juang et al. [8] and van Meirvenne and Goovaerts [2] considered how the proximity of predictions to the threshold could be incorporated into a design algorithm. They suggested that the most beneficial locations to make additional observations are where $|\widehat{z}(\mathbf{x}) - z_c|/\sigma$ is small. Thus they could order every potential observation location according to this criterion. This approach led to clusters where it was desirable to observe the contamination because existing observations were sparse and predictions were close to z_c. However they could not forecast the effect that additional sampling would have on this criterion because the new value of $\widehat{z}(\mathbf{x})$ depended on the new observations. Therefore they had to make intuitive decisions about the intensity with which each cluster was sampled and the total number of observations.

Demougeot-Renard et al. [9] addressed this problem in a multi-phase survey of soil contamination at a former smelter in France. Following the initial survey, they selected additional sampling sites which greatly reduced the cost of misclassifying the remediation requirements of the soil. They simulated an observation, conditional upon the existing observations, at each site in their proposed design. They then used these simulated ob-

servations to estimate the cost and to determine whether the design was fit-for-purpose. However, because their updated objective function was calculated from a single realization of the new design they could not determine the uncertainty associated with it or know if it was truly representative of the proposed design. Also rather than using a numerical algorithm to optimize their additional sampling they compared the values of their objective function for different designs which were selected according to intuitive rules.

We develop a Monte-Carlo multi-phase sampling strategy. Later phases of the survey are optimized to minimize the expected total loss from misclassifications of the contamination status of the soil. The expected total loss is estimated through multiple conditional simulations from a parametric model of spatial variation that is fitted to available data. The expected loss is referred to as the objective function of the optimization and it is minimized by a numerical procedure called spatial simulated annealing (SSA; [10]). Our algorithm is an advance upon existing techniques for the optimization of multiphase surveys since it ensures that the effect of the proposed phase of sampling upon the objective function can be forecast and because the objective function is a direct measure of the effectiveness of the survey rather than an arbitrarily selected measure of the uncertainty or accuracy. Therefore it is possible to optimize simultaneously the locations of multiple observations and to assess whether it is cost-effective to conduct additional phases of different sizes.

The strategy is tested on a survey of soil lead contamination in Glebe, Sydney [11]. Parametric models of spatial variation commonly assume a Gaussian marginal distribution but this is not appropriate in this case since the distribution of the observed lead concentrations is highly skewed. It is known that the misspecification of a spatial model

can cause simulations from it to poorly reflect the actual variation of the observed property [12]. Therefore a logarithmic or Box-Cox transform is often applied to skewed data prior to analysis. In this paper we fit a parametric model of spatial variation to the data within the more general copula framework [13]. Within this framework a range of models with different assumed marginal distributions can be fitted and the quality of fit can be compared according to the Akaike Information Criterion (AIC) [14]. A similar model could have been fitted using a trans-Gaussian kriging framework [15], [16]. The problems of sample design have previously been addressed for copula [17] and trans-Gaussian models [18].

Initially we optimize the survey design to map lead contamination and make remediation decisions at the point-scale. However remediation is generally conducted over larger blocks and the methodology should be up-scaled. For Gaussian properties this up-scaling could be achieved through block kriging [7]. We up-scale the non-Gaussian lead model by averaging multiple point-scale simulations from within each block. We demonstrate that surveys for block-scale recommendations can be achieved by this method although considerable computation time is required.

2. Materials and Methods

2.1 The Glebe survey of soil lead.

The data used in this study were 438 observations of topsoil lead extracted from sites within the Sydney suburb of Glebe in 1993 [19]. Glebe was first established as a residential area in 1828 and by the time of the survey had developed into a high density inner-city

suburb surrounded by major roads and industry [20]. Industrial sites within Glebe and its surrounds have included tanneries, piggeries, abattoirs, jam factories metal foundries, coppersmiths, paint manufacturers and various timber industries [21], [22].

The observation sites in the survey were chosen by a stratified random sampling design (Figure 1). The study area was divided into 227 square cells of 100-m length. One location was randomly selected from the sites with accessible soil within each cell. No observations were collected from eight of the cells where soil was absent. At each selected site, two soil samples were extracted 1 m apart and analyzed separately. The total soil lead content of each sample was determined by flame atomic absorption spectrophotometry on a Varian (Melbourne, Australia) SpectrAA-20 with background correction. Full details of the laboratory procedures are given by Markus and McBratney [20].

These data have previously been analyzed by Cattle et al. [11]. They compared the relative merits of different kriging methods to predict whether the lead concentration at non-sampled sites exceeded the Australian Environmental Investigation Limit (AEIL) of 300 mg kg^{-1}. They found that multiple indicator kriging yielded the most accurate delineation although the copula methodology was not available at that time.

3. Theory

3.1 Non-Gaussian geostatistical models

Conventional geostatistical methods assume that the property of interest is a realization of a second order stationary random variable. A model which describes the spatial correlation of the random function is fitted to the n observed data, $\mathbf{z} = (z_1, z_2, \ldots, z_n)$, where

$z_i = z(\mathbf{x}_i)$ at location $\mathbf{x}_i$. Then this model is used to predict the property across the region by kriging. Kriging yields both a prediction of the property at a particular site and an associated prediction variance referred to as the kriging variance. If the spatial model is fitted by the conventional method of moments [7] then no explicit assumption about the statistical distribution of the random variable is required. However, the method of moments estimator is known to be inefficient if the data are highly skewed [23], and a distributional assumption is required to determine a probability density function (pdf) of the property at each site and to determine the probability that it exceeds a critical threshold.

Model-based geostatistical methods [24] assume a particular distribution for the random variable, most usually a multivariate Gaussian distribution. A function describing the spatial correlation of the distribution is fitted by a likelihood method and when this model is used in the kriging predictor it is referred to as the empirical best linear unbiased predictor (EBLUP; [25]). Since the distribution of the prediction is known, the pdf can be determined and used to calculate the probability that a threshold is exceeded. However the assumption of a multivariate Gaussian distribution is very restrictive and is rarely appropriate for surveys of soil metals around industrial sites where there tends to be a mixture of diffuse underlying pollution and isolated hot spots or outliers.

Bárdossy and Li [26] and Kazianka and Pilz [27] showed that the assumption that a property is a realization of a multivariate Gaussian distribution can be relaxed by use of a copula-based model. In such a model, the marginal distribution and dependence structure are specified separately. It is possible to specify a non-Gaussian dependence structure which permits a different dependence between large and small values. However

such dependence models require intensive computation [26] and are therefore beyond the scope of this study. The specification of a non-Gaussian marginal distribution is itself a marked generalization of the standard Gaussian geostatistical model.

If we denote the distribution function of a property by F the density by f and the Gaussian distribution function with zero mean and unit variance as $\Phi_{0,1}$ and define $\mathbf{a} = \left[\Phi_{0,1}^{-1}\left\{F\left(z_1\right)\right\}, \ldots, \Phi_{0,1}^{-1}\left\{F\left(z_n\right)\right\}\right]$ then for a property with a Gaussian dependence structure, $\mathbf{a}$ is a realization of a multivariate Gaussian random variable and the log-likelihood of the observed data is

$$l\left(\boldsymbol{\theta}\right) = -\frac{1}{2}\log\left|\mathbf{Q}\right| + \frac{1}{2}\mathbf{a}^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{Q}^{-1}\right)\mathbf{a} + \sum_{i=1}^{n}\log\left\{f\left(z_i\right)\right\}. \tag{1}$$

Here $\boldsymbol{\theta}$ is the vector of parameters of both the marginal distribution and the correlation model, $\mathbf{Q}$ is the correlation matrix of $\mathbf{a}$, $\mathbf{I}_n$ is the $n \times n$ identity matrix and T denotes the transpose of a matrix. The elements of the correlation matrix are determined from a parameterized correlation function such as the Matérn function [28]

$$\begin{aligned} Q\left(h\right) &= \left(1-s\right)\left\{\frac{1}{2^{\nu-1}\Gamma\left(\nu\right)}\left(\frac{h}{d}\right)^{\nu}K_{\nu}\left(\frac{h}{d}\right)\right\} \text{ for } h > 0, \\ Q\left(h\right) &= 1 \text{ for } h = 0, \end{aligned} \tag{2}$$

where h is the distance between two observations, s is the proportion of the variance which is spatially uncorrelated, d is a spatial parameter, ν is a smoothness parameter, K_{ν} a modified Bessel function of the second kind of order ν and Γ is the gamma function. The variance of $\mathbf{a}$ is equal to one because the variance of the property is accounted for in the marginal distribution. The correlation function approaches one asymptotically and hence does not have a finite range. We define the effective range d_e, which depends on d and ν, as the lag at which $Q(h) = 0.95(1-s)$.

Thus a copula-based model with any specified distribution function F can be fitted by finding $\widehat{\boldsymbol{\theta}}$ which maximizes Equation (1). The quality of the fit of models with different marginal distributions can be compared by calculation of the AIC [14]:

$$\mathrm{AIC} = 2p - 2l\left(\widehat{\boldsymbol{\theta}}\right), \tag{3}$$

where p is the total number of parameters in the model. The AIC weighs the likelihood against the number of parameters with the smallest value corresponding to the model which has appropriate complexity to describe the variation of the property. Marchant et al. [13] applied copula-based models with Gaussian dependence structures to observations of cadmium across France. They found that a model with a generalized extreme value marginal distribution was a better fit than models with a Gaussian, log-Gaussian or Box-Cox distribution.

Once the most appropriate model has been selected it can be used within a copula kriging algorithm to predict the pdf of the property an any unobserved site. The density of the prediction at a target site is

$$f_t(z_0|\mathbf{z}, \boldsymbol{\theta}) = \frac{f\left(z_0\right)\phi_{\widehat{e}_\mathrm{t},\widehat{v}_\mathrm{t}}\left(a_0\right)}{\phi_{0,1}\left(a_0\right)}, \tag{4}$$

where $a_0 = \Phi_{0,1}^{-1}\left\{F\left(z_0\right)\right\}$, $\phi_{m,v}$ is the Gaussian density function with mean m and variance v, $\widehat{e}_\mathrm{t}$ is the prediction of the expectation of $a = \Phi_{0,1}^{-1}\left\{F\left(z\right)\right\}$ at the target site calculated by simple kriging of $\mathbf{a}$ and $\widehat{v}_\mathrm{t}$ is the corresponding ordinary kriging variance. The expectation and kriging variance of a are calculated from

$$\widehat{e}_\mathrm{t} = \mathbf{Q}_\mathrm{to}\mathbf{Q}^{-1}\mathbf{a}, \tag{5}$$

$$\widehat{v}_\mathrm{t} = \left(1 - \mathbf{Q}_\mathrm{to}\mathbf{Q}^{-1}\mathbf{Q}_\mathrm{ot}\right). \tag{6}$$

Here $\mathbf{Q}_{\mathrm{to}} = \mathbf{Q}_{\mathrm{ot}}^{*}$ denotes the unconditional correlation matrix between $\mathbf{Z}$ at the observation and target locations. The conditional pdf of $z(\mathbf{x}_t)$ can be determined by calculating $f_{\mathrm{t}}(z|\mathbf{z}, \boldsymbol{\theta})$ across the range of plausible z. The distribution function can be determined by numerical integration of the density.

It is possible to generate simulations of z from the copula-based model. Simulations of a at unobserved sites, conditional on the fitted covariance model parameters and the observed $\mathbf{a}$, can be generated by LU simulation [29]. If we denote a realization of spatially correlated values of a at multiple locations by $\mathbf{a}_{\mathrm{s}}$ then the quantiles of these values are $\mathbf{u}_s = \Phi_{0,1}(\mathbf{a}_s)$ and the simulated z are $\mathbf{z}_{\mathrm{s}} = F^{-1}(\mathbf{u}_{\mathrm{s}})$. Full details of the copula methodology and its relation to other geostatistical models are given by Marchant et al. [13].

3.2 Optimization of sampling schemes.

Spatial simulated annealing [10] is a stochastic algorithm which may be used to optimize the configuration of observation locations in a geostatistical survey. If a proposed survey is to consist of n observations then SSA finds the length n vector $\mathbf{X}$ of sampling locations which minimizes a specified objective function $\rho(\mathbf{X})$. The algorithm has been used to minimize various measures of the uncertainty associated with a geostatistical survey (e.g. [30], [31]). The algorithm requires that the objective function can be calculated prior to sampling i.e. it must not be a function of the value of the property of interest at the proposed sample sites.

We note that, in contrast to design based surveys, there is no necessity for observations in geostatistical surveys to be randomly located. This is because in the geostatistical model the assumption of randomness attaches to the realizations of the random function

rather than the sample design [32]. However, biased model estimates can result if the selected local sampling intensity is related to the expected value of the property [33]. Such a situation can arise in a geostatistical survey if, for example, a survey of an ore body is biased towards locations where large concentrations of the ore are expected. Then the observations used to fit the model of variation will be unrepresentative of the true variation. Diggle et al. [33] proposed a model-based strategy to account for such preferential sampling.

4. Calculation

4.1 Case study scenarios

Potential second phases of the Glebe lead survey were optimized to minimize the loss because of misclassifications of remediation requirements. The Glebe survey was used as an illustrative example. In reality, further sampling of the type discussed here would not be appropriate because the initial survey was conducted in 1993 and the soil-lead concentrations might well have changed because of factors such as land use change, soil remediation and natural soil processes. We considered two situations. The first was where a complete list of sites with exposed soil was available and the total loss function from the survey was the sum of the loss function at each of these sites. We denote the vector of locations with exposed soil as $\mathbf{e} = (\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_{n_e})$. Any of the sites with exposed soil could be sampled. For the purpose of this illustrative example we assumed that exposed soil is located on the 952 nodes of a 50-m grid across the study region. The second situation was where the study region was divided into $n_b = 908$ blocks of size 50×50 m

denoted $\mathbf{b} = (\mathbf{b}_1, \mathbf{b}_2, \ldots \mathbf{b}_{n_b})$. The remediation decisions were based upon the mean lead concentration within these blocks. Again any of the 952 exposed sites could be sampled.

4.2 Geostatistical analysis of existing data

Copula-based models with Gaussian dependence structure, Matérn spatial-correlation structure and various marginal distribution functions were fitted to the lead observations by maximum likelihood. Prior to the model fitting the data were scaled such that their variance was one, to reduce the probability of numerical instabilities occurring in the calculation of the log-likelihood. The marginal distributions used were (i) the Gaussian distribution (ii) the log-Gaussian distribution (iii) the Box-Cox distribution and (iv) the generalized extreme value distribution. The formulae for the distribution and density functions are included in the Appendix. The AIC (Equation 3) was calculated for each fitted model and the model with the lowest AIC was used to represent the spatial variation of lead.

The fitted model was used to predict the pdf of lead across the study region by copula kriging (Equation 4). The expected loss from conducting remediation at site e_i

$$L_{\mathrm{R}}\left(e_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) = \int_{z=0}^{z_{\mathrm{c}}} f_{e_i}\left(z|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) L_1\left(z\right) dz, \tag{7}$$

and the expected loss from not conducting remediation

$$L_{\mathrm{N}}\left(e_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) = \int_{z=z_{\mathrm{c}}}^{\infty} f_{e_i}\left(z|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) L_2\left(z\right) dz, \tag{8}$$

was calculated at each prediction site. Here L_1 and L_2 are loss functions for wrongly classifying soil as contaminated and not contaminated respectively.

Cattle et al. [11] suggested such loss functions for the survey of lead in Glebe. One of the false positive loss functions at site $\mathbf{x}$ was

$$L_1 = z_{\mathrm{c}} - z\left(x\right), \tag{9}$$

and the corresponding false negative loss function was

$$L_2 = \alpha\left\{z\left(\mathbf{x}\right) - z_{\mathrm{c}}\right\}, \tag{10}$$

where z_{c} was the AEIL and α a factor which weighs human health costs of false negatives against the unnecessary remediation costs of false positives. Both of these loss functions increased with the magnitude of the misclassification and the α was greater than one to ensure that the loss from false negatives exceeded that from false positives.

Remediation is conducted at a site if and only if $L_{\mathrm{R}} < L_{\mathrm{N}}$. The expected total loss from the entire remediation program conditional on the available observations was

$$L_{\mathrm{T}} = \sum_{i=1}^{n_e} \min\left\{L_{\mathrm{R}}\left(e_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right), L_{\mathrm{N}}\left(e_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right)\right\}. \tag{11}$$

Li et al. [16] consider how decisions can be made in terms of more general utility functions.

When the remediation decisions were made across blocks the loss functions were estimated using 100 conditional realizations of z at 25 sites on a regular grid within each block. The realizations were simulated by the LU method and we denote the simulated value within realization r at site j as $z_{\mathrm{b}}\left(i, j\right)$, $r = 1, \ldots, 100$ and $j = 1, \ldots, 25$. Then

$$L_{\mathrm{R}}\left(b_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) = \frac{1}{100}\sum_{r=1}^{100} L_1\left\{\frac{1}{25}\sum_{j=1}^{25} z_{\mathrm{b}}\left(r, j\right)\right\}, \tag{12}$$

and

$$L_{\mathrm{N}}\left(b_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) = \frac{1}{100}\sum_{r=1}^{100} L_2\left\{\frac{1}{25}\sum_{j=1}^{25} z_{\mathrm{b}}\left(r, j\right)\right\}. \tag{13}$$

4.3 Optimization of a second phase of sampling.

The aim of a second phase of sampling is to reduce the expected loss from the remediation programme as efficiently as possible and to ensure that the reduction in this loss exceeds the cost of additional sampling. If the additional phase consists of n observations located at sites $\mathbf{s} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n)$ then the density of z at each exposed site, and hence the loss functions L_{R}, L_{N} and L_{T} are conditional on $\mathbf{s}$ in addition to $\mathbf{z}$ and $\widehat{\boldsymbol{\theta}}$. One might expect to re-estimate $\widehat{\boldsymbol{\theta}}$ subsequent to the additional sampling but the design of this sampling is guided from the results of the first phase and hence depends on $\mathbf{z}$ so a re-estimate of $\widehat{\boldsymbol{\theta}}$ will be biased. In the Glebe survey the uncertainty associated with the $\widehat{\boldsymbol{\theta}}$ from the initial survey should be small since it is based on more than 400 observations with more than 200 pairs separated by 1 m [34]. The expected loss function subsequent to additional sampling is

$$\mathrm{E}\left\{L_{\mathrm{T}}\left(\mathbf{s}\right)\right\} = \sum_{i=1}^{n_e} \min\left\{\int_{\mathbf{Z}(\mathbf{s})} L_{\mathrm{R}}\left(e_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) d\mathbf{z}\left(s\right), \int_{\mathbf{Z}(\mathbf{s})} L_{\mathrm{N}}\left(e_i|\mathbf{z}, \widehat{\boldsymbol{\theta}}\right) d\mathbf{z}\left(\mathbf{s}\right)\right\}, \tag{14}$$

where $\mathbf{Z}\left(\mathbf{s}\right)$ denotes the complete space of realizations of the random variable, conditional upon the existing observations and at the proposed sampling locations. We approximate $\mathbf{Z}\left(\mathbf{s}\right)$ by $n_{\mathrm{sim}} = 1000$ realizations of $\mathbf{z}\left(\mathbf{s}\right)$ generated by conditional LU simulation [30]. If each of these realizations is denoted $\mathbf{z}\left(\mathbf{s}\right)_r$ then the total loss function becomes

$$\mathrm{E}\left\{L_{\mathrm{T}}\left(\mathbf{s}\right)\right\} = \frac{1}{n_{\mathrm{E}}} \sum_{i=1}^{n_e} \min\left\{\sum_{j=1}^{n_{\mathrm{sim}}} L_{\mathrm{R}}\left(e_i|\mathbf{z}\left(\mathbf{s}\right)_j, \mathbf{z}, \widehat{\boldsymbol{\theta}}\right), \sum_{j=1}^{n_{\mathrm{sim}}} L_{\mathrm{N}}\left(e_i|\mathbf{z}\left(\mathbf{s}\right)_j, \mathbf{z}, \widehat{\boldsymbol{\theta}}\right)\right\}. \tag{15}$$

We optimize the locations $\mathbf{s}$ of the observations in a new phase of sampling by SSA with objective function $\rho\left(\mathbf{s}\right) = \mathrm{E}\left\{L_{\mathrm{T}}\left(\mathbf{s}\right)\right\}$. This procedure was used to optimize second phase surveys of 10, 20, 30, 40, 50 observations when remediation decisions were made at

the point scale. The additional phases of sampling were initially optimized for the loss functions (Equations 9-10) suggested by Cattle et al. [11]. The exercise was then repeated with loss functions of the same form but the critical threshold increased to 1300 mg kg^{-1} to illustrate how the optimal schemes change as the proportion of the study region in need of remediation changes. There was a substantial increase in the computation time required when the remediation decisions were made across blocks. Therefore only one illustrative second phase of 30 points was designed.

5. Results

The histogram of observed lead concentrations in Glebe (Fig 1b) included extreme values and was highly skewed (skew=6.44) and hence the Gaussian function was not suitable to describe the marginal distribution. The fitted model with a Gaussian marginal had the smallest log-likelihood and largest AIC of the four candidate models (Table 1). The model with a Box-Cox marginal distribution had the largest log-likelihood and the smallest AIC and was therefore used to predict the lead content across the study area. For this fitted model, spatial correlation is evident up to an effective range of 234 m. The map of the expected lead concentration (Fig. 2a) is dominated by one hotspot on the western boundary where concentrations were almost 12 000 mg kg^{-1}. The probability that the AEIL threshold of 300 mg kg^{-1} is exceeded at this location is close to 1. The probability of exceeding this threshold is greater than 0.8 for 12 % of sites in the study region. These sites are generally located in the centre of the study area. For 12 % of sites, mostly located on the northern, eastern and southern boundaries, the probability that the threshold is exceeded is less than 0.2. The map of the loss function upon remediation (Fig. 3b) is

roughly the inverse of the probability map.

The expected lead concentration is greater than the AEIL at 74 % of sites and the probability that the AEIL is exceeded is greater than 0.5 at 53% of sites. The Monte Carlo uncertainty analysis suggests that remediation should be conducted for 90 % of the study region (Fig 4a) and the expected loss is 66.9 monetary units per site (Table 2). If the AEIL were raised to 1300 mg kg^{-1} remediation would only be cost-effective at 44% of sites (Fig 4c) with expected loss of 57.5 monetary units per site. The expected concentration exceeds this modified threshold at 17 % of sites and the probability of exceedance is greater than 0.5 at 2 % of sites.

Figures 4a and 4b show the optimized locations of 30 observations in second phase surveys where the thresholds are 300 mg kg^{-1} and 1300 mg kg^{-1} respectively. Figure 5 shows a 30 point optimized design where remediation requirements are assessed over 50 m blocks rather than at the point scale. In all designs, observations are concentrated upon the boundaries between where remediation is required and not required. For the point scale surveys the 30 point second phase surveys reduced the expected loss to 63.5 monetary units when the threshold was 300 mg kg^{-1} and to 51.9 monetary units when the threshold was 1300 mg kg^{-1} (Table 2). If the costs of sampling are known then this information can be used to decide whether a second phase of sampling is cost-effective and to determine the optimal size of this second phase.

6. Discussion and Conclusions

The copula framework was used to assess the relative suitability of various models of the variation of lead around Glebe. The parametric form of the model meant that condi-

tional simulations of lead could be easily generated. Hence it was possible to conduct an uncertainty analysis of lead predictions, conditional on the observed data, and to account for the whole pdf when deciding whether remediation is required at a certain site. This decision was based on a comparison of the expected losses from remediating and not remediating.

Through the efficient use of LU simulation and copula-kriging it was also possible to forecast the expected loss functions which would result from a proposed additional phase of sampling. By comparison of the forecast reduction in the loss with the costs of the extra sampling, an informed decision could be made about whether to conduct the extra sampling. The forecast loss function was used in an SSA algorithm to optimize the locations of observations within additional phases of the survey. These optimized surveys located observations at the boundaries between areas which require and do not require remediation. If additional phases of different sizes are optimized it is possible to select the optimal number of observations. We note that for both thresholds considered, the areas where remediation was recommended were substantially larger than the regions where the probability of exceeding the threshold was greater than 0.5. This suggests that locating additional observations at sites where the probability of exceedance is close to 0.5 is a sub-optimal sampling approach.

The methodology described here is applicable for any specified loss function. However it does rely upon the availability of such a loss function. The loss function used in this study was largely illustrative but more realistic functions have been used in other studies. Ramsay et al. [35] developed loss functions for a number of specific contamination sites to use within their *optimized contaminated land investigation methodology.*

Demougeot-Renard et al. [9] derived a loss function for a survey of a former smelting works. Brus et al. [36] showed how mathematical models can be used to quantify the impact of soil metal pollution on crops and cattle and form the basis of loss functions.

The model of variation of the soil lead content is fitted to only the initial phases of sampling. Updating the model after each phase of sampling could lead to bias because areas where lead concentrations are expected to be close to the critical threshold are over-represented in the additional phases. This was unlikely to be an issue with the Glebe survey because the initial sampling consisted of more than 400 observations and a large number of comparisons over short distances and was hence very suitable for geostatistical model fitting [34]. However in circumstances where the initial model is too uncertain it would be possible to account for the preferential sampling in the later phases [33] or to design phases of sampling which do not depend on $z(\mathbf{x})$ specifically to reduce the model uncertainty [3]. Also an objective function which accounts for both variogram uncertainty and the kriging variance could be used [34], [37]. Such a objective function tends to lead to close pairs of sampling locations within the survey.

The methodology fits a model of variation to the observed data so it accounts for both the uncertainty because of sampling and the uncertainty because of the laboratory analysis of the samples. If multiple samples from the same site are analyzed then it is possible to separate these components of uncertainty in the model (e.g. [38]). Then the SSA approach could be used to explore the potential benefits of repeated analyses on soil from the same site to reduce the analytical uncertainty.

From a geostatistical perspective the methodology is novel because it accounts for both the uncertainty and expected concentration at each site and it can forecast the effect

of multiple potential observations. The benefit from a practical perspective are that it can calculate the expected loss function from existing observations and forecast what the loss function will be if further phases of sampling are conducted. Hence it can be used to make an informed decision about optimal remediation strategies and further sampling.

Acknowledgements

BPM and RML were funded by the Biotechnology and Biological Sciences Research Council through its Institute Strategic Programme Grant to Rothamsted Research. BPM is grateful to The University of Sydney for the award of an International Visiting Research Fellowship. His contribution was conducted during his visit to the Faculty of Agriculture, Food and Natural Resources at the University of Sydney. RML's contribution is published with the permission of the Director of the British Geological Survey (NERC).

References

[1] Environment Agency, Guidance on the use of soil screening values in ecological risk assessment, Environment Agency Report SC050021, Environment Agency, Bristol, 2008, pp. 37.

[2] M. van Meirvenne, P. Goovaerts, Evaluating the probability of exceeding a site-specific soil cadmium contamination threshold, Geoderma 102 (2001) 75–100.

[3] B.P. Marchant, R.M. Lark, Adaptive sampling for reconnaissance surveys for geostatistical mapping of the soil, European Journal of Soil Science 57 (2006) 831–845.

[4] S. Verstraete, M. van Meirvenne, A multi-stage sampling strategy for the delineation

of soil contamination in a contaminated brownfield, Environmental Pollution 154 (2008) 184–191.

[5] E. Meerschman, L. Cockx, M. van Meirvenne, A geostatistical two-phase sampling strategy to map soil heavy metal concentrations in a former war zone, European Journal of Soil Science 62 (2011) 408–416.

[6] G.B.M. Heuvelink, Z. Jiang, S. de Bruin, C.J.W Twenhöfel, Optimization of mobile radioactivity networks, International Journal of Geographical Information Science 24 (2010) 365–382.

[7] R. Webster, M.A. Oliver, Geostatistics for Environmental Scientists, second ed., John Wiley and Sons, Chichester, 2007.

[8] K.W. Juang, W.-J. Liao, T.L. Liu, L. Tsui, D.-Y. Lee, Additional sampling based on regulation threshold and kriging variance to reduce the probability of false delineation in a contaminated site, Science of the Total Environment 389 (2008) 20–28.

[9] H. Demougeot-Renard, C. de Fouquet, P. Renard, Forecasting the number of soil samples required to reduce remediation cost uncertainty, Journal of Environmental Quality 33 (2004) 1694–1702.

[10] J.W. van Groenigen, G. Pieters, A. Stein, Constrained optimization of soil sampling for minimisation of the kriging variance, Geoderma 87 (1999) 239–259.

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

[12] C. Hofer, A. Papritz, Predicting threshold exceedance by local block means in soil pollution surveys, Mathematical Geosciences 42 (2010) 631–656.

[13] B.P. Marchant, N.P.A. Saby, C.C. Jolivet, D. Arrouays, R.M. Lark, Spatial prediction of soil properties with copulas, Geoderma 162 (2011) 327–334.

[14] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: B.N. Petrov, F. Csáki (Eds.), Second International Symposium on Information Theory, Akadémiai Kiadó, Budapest, 1973, pp. 267–281.

[15] M.O. Prates, D.K. Dey, M.R. Willig, J. Yan, Transformed Gaussian Markov random fields and spatial modelling, Technical report, University of Connecticut, Statistics Department (2012) http://arxiv.org/abs/1205.5467.

[16] V.D. Oliveira, B. Kadem, D.A. Short, Bayesian prediction of transformed Gaussian random fields, Journal of American Statistical Association 92 (1997) 1422–1497.

[17] J. Li, A. Bárdossy, L. Guenni, M. Liu, A copula based opservation network design approach, Environmental Modelling and Software 26 (2011) 1349–1357.

[18] J. Pilz, H. Kazianka, G. Spöck, Some advances in Bayesian spatial prediction and sampling design, Spatial Statistics 1 (2012) 65–81.

[19] J.A. Markus, A survey of heavy metals in the topsoil of Glebe, B.Sc. Honours thesis, Dep. of Aric. Chem. and Soil Sci., Univ of Sydney, Australia, 1993.

[20] J.A. Markus, A.B. McBratney, An urban soil study: Heavy metals in Glebe, Australia, Australian Journal of Soil Research 34 (1996) 453–456.

[21] J.R. Sands, Sands' Sydney and Suburban Directory, John Sands, Sydney, 1899.

[22] F. MacDonnell, Local industry, in: The Glebe: Portraits and Places, Ure Smith, Sydney, 1975, pp. 75–84.

[23] R.M. Lark, A comparison of some robust estimators of the variogram for use in soil survey, European Journal of Soil Science 51 (2000) 137–157.

[24] P.J. Diggle, P.J. Ribeiro Jr., Model-based Geostatistics, Springer, New York, 2007.

[25] R.M. Lark, B.R. Cullis, S.J. Welham, On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML, European Journal of Soil Science 57 (2006) 787–799.

[26] A. Bárdossy, J. Li, Geostatistical interpolation using copulas, Water Resources Research 44 (2008) W07412.

[27] H. Kazianka, J. Pilz, Copula-based geostatistical modeling of continuous and discrete data including covariates, Stochastic Environmental Research and Risk Assessment 24 (2010) 661–673.

[28] B. Matérn, Spatial variation, Meddelanden från Statens Skogsforskningsinstitut, 49 (1960) No. 5. [Second ed., Lecture Notes in Statistics, No. 36, Springer, New York, 1986].

[29] C.V Deutsch, A.G. Journel, GSLIB: Geostatistical Software Library and User's Guide, second ed., Oxford University Press, New York, 1998, pp.146–147.

[30] D.J. Brus, G.B.M. Heuvelink, Optimization of sample patterns for universal kriging of environmental variables, Geoderma 138 (2007) 86–95.

[31] S.J. Melles, G.B.H. Heuvelink, C.J.W. Twenhöfel, A. van Dijk, P.H. Hiemstra, O. Baume, U. Stohlker, Optimizing the spatial pattern of networks for monitoring radioactive releases, Computers and Geosciences 37 (2011) 280–288.

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

[33] P.J. Diggle, R. Menezes, S. Ting-li, Geostatistical inference under preferential sampling, Journal of the Royal Statistical Society, C 59 (2010) 191–232.

[34] B.P. Marchant, R.M. Lark, Optimal sampling for geostatistical surveys, Mathematical Geology 39 (2007) 113–134.

[35] M.H. Ramsey, P.D. Taylor, J. Lee, Optimized contaminated land investigation at minimum overall cost to achieve fitness-for-purpose, Journal of Environmental Monitoring 4 (2002) 809–814.

[36] D.J. Brus, J.J. de Gruijter, D.J.J. Walvoort, F. de Vries, J.J.B. Bronswijk, P.F.A.M Romkens, W. de Vries, Mapping the probability of exceeding critical thresholds for cadmium concentrations in soils in the Netherlands, Journal of Environmental Quality 31 (2002) 1875–1884.

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

[38] B.P. Marchant, A.M. Tye, B.R. Rawlins, The assessment of point-source and diffuse soil metal pollution using robust geostatistical methods: a case study in Swansea (Wales, UK), European Journal of Soil Science 62 (2011) 346–358.

Appendix

Distribution and density functions

The formulae for marginal distribution and density functions considered in this paper are:

Gaussian distribution

$$F^{\mathrm{G}}(z) = \Phi_{\mu,\sigma^2} = \frac{1}{2}\left\{1 + \mathrm{erf}\left(\frac{z-\mu}{\sigma\sqrt{2}}\right)\right\},\ f^{\mathrm{G}}(z) = \phi_{\mu,\sigma^2} = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{(z-\mu)^2}{2\sigma^2}\right\}, \tag{16}$$

where erf is the error function.

Log-Gaussian distribution

$$F^{\mathrm{L}} = F^{G}(z^*),\ f^{\mathrm{L}} = \frac{f^{\mathrm{G}}(z^*)}{z}, \tag{17}$$

where $z^* = \log(z)$.

Box-Cox Gaussian distribution

$$F^{\mathrm{B}} = F^{G}(z^*),\ f^{\mathrm{B}} = \frac{f^{\mathrm{G}}(z^*)}{z^{1-\lambda}}, \tag{18}$$

where $z^* = (z^\lambda - 1)/\lambda$.

Generalized Extreme Value distribution

$$F^{E} = \exp\left(-T^{-\frac{1}{\xi}}\right),\ f^{E} = \frac{1}{\sigma}\left(T^{-\frac{1}{\xi}-1}\right)\exp\left(-T^{-\frac{1}{\xi}}\right), \tag{19}$$

where μ is the location parameter, σ the scale parameter, ξ the shape parameter, $T = 1 + \xi(z-\mu)/\sigma$ and the distribution exists for $T > 0$.

Table 1: Fitted variogram parameters in scaled units, likelihoods and AIC values for different marginal distributions.

	Gaussian	Log-Gaussian	Box-Cox	GEV
s	0.00	0.00	0.00	0.00
d m	712	119	114	615
ν	0.19	0.22	0.21	0.20
d_e	1404	250	234	1231
μ	-0.11	-1.89	-2.00	-0.07
σ	0.34	1.36	1.40	0.99
λ	-	-	-0.04	-
ξ	-	-	-	0.01
L	359.1	406.0	408.2	394.7
AIC	-708.2	-802.0	-804.4	-777.5

Table 2: Expected losses from misclassications after optimized second phase survey of size N for different threshold values.

N	$z_c = 300$	$z_c = 1300$
0	66.8700	57.5300
10	65.2139	55.3182
20	64.2117	53.5228
30	63.5089	51.8942
40	62.8680	50.4714
50	62.4028	48.9964

Figure Captions

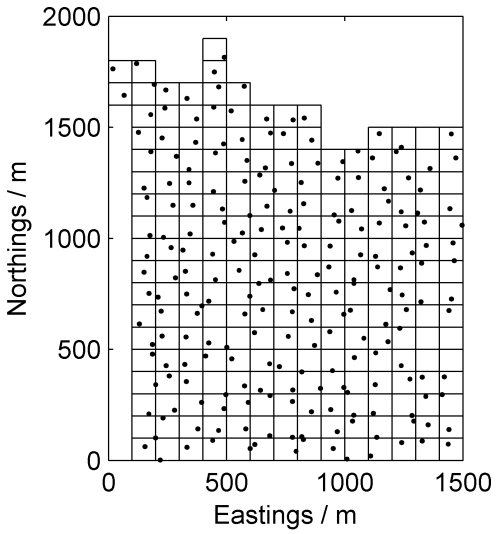
Figure 1: Sample scheme for 1993 survey of soil lead in Glebe, Australia with 100-m grid used for stratification superimposed (left) and histogram of lead observations (right).

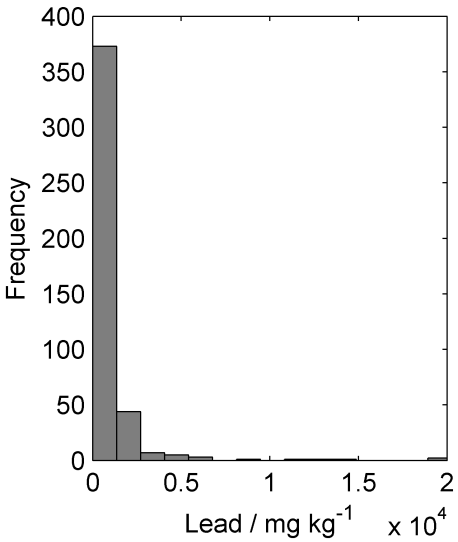
Figure 2: (left) Spatial prediction of expected concentration of soil lead in Glebe from 1993 survey. (right) Spatial prediction of the probability that the soil lead concentration exceeds the AEIL Regulatory threshold of 300 mg kg^{-1}.

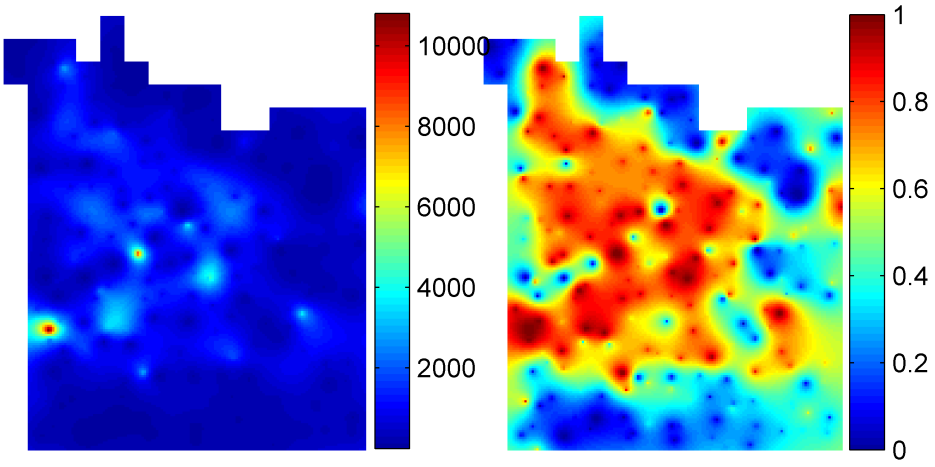
Figure 3: Expected loss functions for 1993 Glebe survey if soil is (left) classified as not contaminated or (right) classified as contaminated.

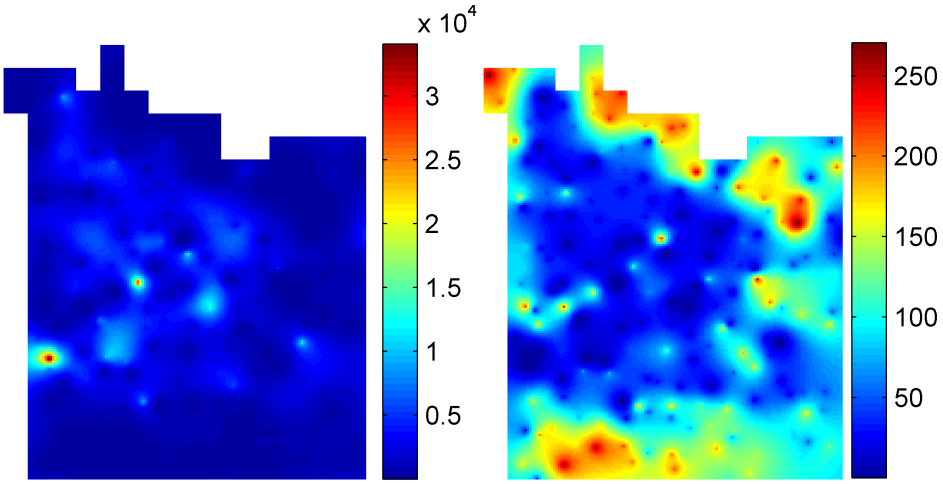
Figure 4: Optimized 30 observation second phase sample schemes superimposed upon point-scale remediation recommendation from the 1993 Glebe survey. Remediation is recommended for black regions and not for grey ones. Critical thresholds are 300 mg kg^{-1} (left) and 1300 mg kg^{-1} (right).

Figure 5: Optimized 30 observation second phase sample schemes superimposed upon block remediation recommendation from the 1993 Glebe survey. Remediation is recommended for black regions and not for grey ones. Critical thresholds is 300 mg kg^{-1}.


2000
1500
1000
500
0
Northings / m
0
500
1000
1500
Eastings / m


400
350
300
250
200
150
100
50
0
Frequency
0
0.5
1
1.5
2
Lead / mg kg-1
x 10^4


10000
8000
6000
4000
2000
1
0.8
0.6
0.4
0.2
0


x 10^4
3
2.5
2
1.5
1
0.5
250
200
150
100
50