Can we predict the provenance of a soil sample for forensic purposes by reference to a spatial database?

R.M. Lark[1] & B.G. Rawlins[2]

[1]*Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK* and [2] *British Geological Survey, Keyworth, Nottingham NG12 5GG, UK*

Short title: *Predicting the provenance of a soil sample*

Correspondence: R.M. Lark. E-mail: murray.lark@bbsrc.ac.uk

Summary

In forensic soil science it is sometimes necessary to address a question of the form: 'what is the most likely place of origin of this soil material', where the possible provenances are in a large area. This 'intelligence' problem may be distinguished from the 'evidence' problem where we need to evaluate the grounds for believing that some soil material is derived from one site rather than another. There is interest in the use of soil databases to solve intelligence problems. This paper proposes a geostatistical method to tackle the intelligence problem. Given data on a sample of unknown provenance, and a database with the same information from known sites, it is possible to define a likelihood function, the argument of which is location in space, which is the likelihood that the sample is from that location. In this paper we show how an approximation to this likelihood can be computed, using a principal component transformation of the data and disjunctive kriging.

The proposed likelihood function is tested using a geochemical database on the soil of the Humber Trent region of north-east England. This shows that the function is a useful way to make a statistical prediction of the provenance of a soil sample. The region can be stratified according to the value of the likelihood function. A validation data set showed that if we defined a stratum with the top 4.5% of values of the likelihood function, then there was a 50% probablity that it included the true provenance of the sample, and there is a 90% probability of finding the true provenance of the sample in a stratum with the top 30% of values of the likelihood function. Note also that the spatial likelihood function could be integrated with other sources of information on the likely provenance of the sample by means of Bayes law.

We conclude that this approach has value for forensic problems. The main difficulty is how to define the geostatistical support of the forensic specimen, and the reliability of analytical data on relatively small forensic samples, but this is a generic problem for forensic geoscience.

Introduction

'You have come up from the south-west, I see.'

'Yes, from Horsham.'

'That clay and chalk mixture which I see upon your toe caps is quite distinctive'.

Sir Arthur Conan Doyle, *The Five Orange Pips*.

Since most of the earth's surface is covered by soil it is not surprising that there is a long-standing interest in the use of soil evidence for forensic purposes (Pye, 2007). In this paper we consider the case when soil has been found on a vehicle, a tool or some other exhibit, and the aim of the forensic investigator is to identify the likely provenance of this soil, or to exclude potential provenance regions from an investigation. Since the soil is very variable at all spatial scales neither the matching of a specimen to a provenance, nor the exclusion of possible provenances, which forensic scientists often wish to achieve, can usually be absolute. For this reason we set the task of inferring forensic intelligence from soil data in a statistical framework. This may be problematic for presenting evidence in court, but may be useful for forensic intelligence (i.e. as a guide to police during an investigation).

Forensic scientists have to make inferences of this kind for real problems. For example, in 2000 three people were reported missing in South Australia. Their vehicle was later recovered; a shovel was in the boot with a lot of soil on it. Examination of the soil, its chemical properties, lithology, mineralogy and organic status, allowed soil scientists to narrow down its likely provenance, and this led directly to the discovery of the remains of the missing persons in a quarry (CAFSS, 2006).

The problem of how best to determine the provenance of soil material for forensic purposes is a matter of considerable interest. For example, Rawlins *et al.* (2006) conducted a study in which four experts used different technologies (X-ray diffraction, scanning

electron microscopy, palynology and molecular characterization of organic matter) to examine soil specimens from different settings and identify their likely provenance. Two sites, with distinctive vegetation and parent material, were easily characterized, but a third was not.

An alternative approach would be to compare forensic specimens with existing soil databases. In the UK, and in many European countries, there are substantial databases on the soil which have been collected to characterize soil resources, and as a baseline to monitor their quality (e.g. McGrath & Loveland, 1992). This raises the question of whether a comparison of soil material from a forensic exhibit with soil in such a database would allow the provenance of the forensic specimen to be narrowed down to a useful degree. Soil scientists have made substantial use of such databases to undertake classical geostatistical inference about the soil (spatial prediction). In such inference we start with a body of data on the soil at known discrete sample points, and proceed to predict soil properties at unsampled points. Locations are given at which the values of the soil properties are unknown. If these unsampled points consitute a grid, then the predictions can be used to produce an isarithmic map of soil properties (Burgess & Webster, 1980). In addition we can make other inferences at unsampled sites, we might compute the probability, conditional on our data, that the true value of a variable at some site exceeds a regulatory threshold. This can be done by disjunctive kriging (Matheron, 1976).

The inference for forensic intelligence is rather different. Again, we have the database, on which our inference will be conditioned, but this time we know the values of key soil variables from a specimen of unknown provenance. What we want to do is to predict that provenance, or alternatively, to exclude sites of potential provenance, as a guide to investigators. We use 'prediction' in this paper in a statistical sense. A prediction of an unknown variable is value that is inferred, conditional on some data and some statistical model, that is 'best' by some appropriate criterion. Such a prediction has an attendant uncertainty, which may be quantified, and should not be treated as an unconditional

statement of fact. It is our contention in this paper that a geostatistical inference of the provenance of a soil sample, conditional on a spatial database, is possible via a spatial likelihood function. We present this likelihood function, and suggest how it might be approximated in practice. We then use an existing geochemical database on the soil to illustrate how the spatial likelihood function might be applied.

Theory

The concept of a spatial likelihood function.

Let $\mathbf{S}$ be a random variate (e.g. a set of geochemical soil properties), let $\mathbf{s}(\mathbf{X})$ be a set of observations of this variate at locations in $\mathbf{X}$, let $\mathbf{x}_0$ be some unsampled location and let $\boldsymbol{\phi}$ be a vector of (cross) covariance parameters for $\mathbf{S}$. These parameters may be estimated from the data $\mathbf{s}(\mathbf{X})$.

The conditional probability density function (pdf) for $\mathbf{S}$ at $\mathbf{x}_0$ is

$$\mathcal{P}\left\{\mathbf{S} \mid \mathbf{s}(\mathbf{X}), \mathbf{x}_0, \boldsymbol{\phi}\right\}. \tag{1}$$

Now, if we consider an observed variate, of unknown provenance, $\mathbf{s}'$ we could evaluate the probability density at any $\mathbf{x}_0$ given the data and covariance model. If we think of this conditional probability density as a function of location, conditional on a particular observation, it is a likelihood function:

$$\mathcal{L}\left\{\mathbf{x}_0 | \mathbf{s}(\mathbf{X}), \mathbf{s}', \boldsymbol{\phi}\right\} = \mathcal{P}\left\{\mathbf{s}' \mid \mathbf{s}(\mathbf{X}), \mathbf{x}_0, \boldsymbol{\phi}\right\}. \tag{2}$$

If we evaluated this likelihood function over a grid of locations it could be used to make inferences about the provenance of the soil sample. For example, a prediction of its provenance might be the location where the likelihood is largest. Alternatively, we might integrate the spatial likelihood function with a spatial prior probability density function for the provenance of the sample (which might reflect other evidence which is available) and then renormalize the result to obtain a spatial posterior probability density function. Again, while a prediction of the provenance could be obtained as the site where the

posterior probability density is largest, a map of the posterior probability density will be of most value for intelligence purposes, indicating those regions where searches or other investigations should be focussed. However, the evaluation of the spatial likelihood function is not a trivial task, and we now consider how it could be done in practice.

Distributional assumptions.

In geostatistical prediction we do not generally evaluate the full conditional pdf at some location $\mathbf{x}_0$. Rather we estimate the best linear unbiased predictor, which is the mean of the conditional pdf if the spatial (cross) covariance is known correctly. In the simple cokriging case (where the expectation of $\mathbf{S}$ is assumed to be known and constant, the cross-covariances of the kriging estimates are the covariance matrix of the conditional pdf, so, subject to assumptions of normality, the conditional pdf could be specified.

In practice we do not proceed in this way when conditional probabilities for a random variable are required at unsampled locations, since our conclusions will be sensitive to the distributional assumption which is often not plausible. This is why the methods of non-linear geostatistics have been developed (Rivoirard, 1994). These entail simple kriging prediction of non-linear transforms of the data, such as the indicator transform (indicator kriging) or a Hermite transform (disjunctive kriging). We now consider the latter in more detail.

Disjunctive kriging (DK) entails the assumption that our data are a realization of a process with a second-order stationary bivariate distribution. The assumption of second-order stationarity means that the covariance function exists and that the variogram is therefore bounded. It is also assumed that the data are from a Gaussian random process. Since data may often not resemble a Gaussian random variable the first step in DK is to transform the data with Hermite polynomials, which Rivoirard (1994) describes in more detail. The Hermite coefficients are then kriged to target locations of interest. A prediction of the original soil variable, $\tilde{S}(\mathbf{x})$, is obtained from these and the conditional probability that $S(\mathbf{x})$ occurs in specified intervals. Here we assume that the range of

a variable is divided into bins, and denote by $\psi_k(\mathbf{x}_0)$ the probability that $S(\mathbf{x}_0)$ is in the kth bin. Note that if we develop this approach, we obtain not a probability density function, but rather probabilities for discrete intervals of the variables. This is one sense in which the proposal developed in this paper provides us with an *approximation* to a spatial likelihood function.

The problem of many variables.

We have described disjunctive kriging above with respect to predicting a single variable, but in a forensic context we will probably want to evaluate a spatial likelihood function that is based on a random variate which represents several soil properties. Disjunctive cokriging is possible (e.g. Finke & Stein, 1994). However, all cokriging requires that we can model the spatial covariation of a variate in terms of an admissible model such as the linear model of coregionalization, LMCR (Journel & Huijbregts, 1978). While the LMCR can be fitted automatically (Lark & Papritz, 2003) which means that it is feasible to fit it for variates with many dimensions, it does impose strong assumptions of linearity, and as a result the fitted covariance matrices for the nested components of the model (coregionalization matrices) may often be positive semi-definite only, which represents the best admissible solution, but clearly implies some 'strain' in the fit of the model (since it implies that some of the variates are perfectly correlated). We would therefore prefer to avoid cokriging techniques that require these constraints.

A proposal.

We therefore propose the following approach.

First, we transform our m-variate data set, $\mathbf{s}(\mathbf{X}), = \{\mathbf{s}(\mathbf{x}_1), \mathbf{s}(\mathbf{x}_2), \ldots\}$, to its m principal components, which we denote by $\mathbf{a}(\mathbf{X}), = \{\mathbf{a}(\mathbf{x}_1), \mathbf{a}(\mathbf{x}_2), \ldots\}$. We propose that the principal components analysis (PCA) is based on the sample correlation matrix of $\mathbf{s}(\mathbf{X})$ so that the transform is independent of the units in which the original variables are expressed. Any new vector, $\mathbf{s}'(\mathbf{x}_0)$ can then be transformed to $\mathbf{a}'(\mathbf{x}_0)$, a projection of the vector onto the same rotation of the original variables computed from the correlation ma-

trix of the data $\mathbf{s}(\mathbf{X})$. The principal components are uncorrelated, and so we will assume that they are a realization of m mutually independent random variables, $A_1,\ A_2,\ \ldots, A_m$. By reference to the eigenvalues from the PCA we can identify how many of the principal components are needed to represent some adequate proportion of the variation of $\mathbf{s}(\mathbf{X})$, we assume that $m' \leq m$ are selected.

The next step is to undertake DK estimation of the m' selected principal components at a set of target sites. We divide the range of values of each component into intervals, so by DK we can estimate for any unsampled site, $\mathbf{x}_0$, a set of probabilities: $\psi_{i,k}(\mathbf{x}_0),\ i = 1, 2, \ldots, m'; k = 1, 2, \ldots, K_i$, where $\psi_{i,k}(\mathbf{x}_0)$ denotes the probability that $A_i(\mathbf{x}_0)$ is in the kth inteval for the ith principal component out of K_i such intervals. Note that the intervals are non-overlapping and cover the full range of values so that, for any i, $\sum_{k=1}^{K_i} \psi_{i,k}(\mathbf{x}_0) = 1$.

We now consider a sampled variate, $\mathbf{s}'$ of unknown provenance. First, we transform it to a vector in the principal component space, $\mathbf{a}'$. For each of the m' principal components we can then identify the interval in which the corresponding variable in $\mathbf{a}'$ falls, we denote this interval by the index $\hat{k}$. On the assumption that the random variables, $A_1,\ A_2,\ \ldots, A_{m'}$ are mutually independent, the approximate spatial likelihood function at $\mathbf{x}_0$ for observed variate $\mathbf{s}'$ is then computed as:

$$\breve{\mathcal{L}}(\mathbf{x}_0|\mathbf{s}') = \prod_{i=1}^{m'} \psi_{i,\hat{k}}(\mathbf{x}_0), \tag{3}$$

where the dependence of this likelihood on the data, on the covariance models used to compute the DK estimates and on the selected principal components is implicit.

Case Study.

In the case study we used soil geochemical data from the G-BASE project from across the Humber-Trent region of north-east England. These data were collected and are maintained by the British Geological Survey, and a large proportion of these data have been described in detail elsewhere (Rawlins *et al.*, 2003). In a previous study Rawlins &

Cave (2004) used them to study geochemical variability of soils, and their implications for forensic problems. In summary, the data were obtained by a non-aligned sampling scheme. The strata were 2-km squares of the Ordnance Survey grid. Every second square was sampled at a random location. This gave 6411 sites in total. At each site five soil cores were collected from the centre and corners of a 20-m square; they were then bulked. The cores were 15 cm long and excluded surface litter. The bulked material was air-dried then seived and a 50-g subsample was ground. The total concentrations of 24 major and trace elements were determined in each sub-sample by wavelength dispersive XRFS (X-Ray Fluorescence Spectrometry).

Rawlins & Cave (2004) concluded that, of the 24 elements determined on these sub-samples, 6 were not suitable for further analysis since many of the observations were below the detection limit of the XRFS system. We followed them in using the following 18 determinations for our analysis, with major elements expressed as weight percent of their oxide: As, Ba, CaO, Co, Cr, Cu, Fe_2O_3, MgO, MnO, Mo, Pb, Rb, Sr, TiO_2, U, V, Zr.

We removed 1000 observations from the data set by simple random sampling. These were for later use as a validation subset. We then computed a principal components analysis of the correlation matrix of the remaining prediction data. Figure 1 shows a plot of the accumulated eigenvalues. We selected the first 7 principal components for further analysis, these account for 80% of the variation of the full data set between them.

We then found a Hermite transformation of each of the principal components to a new normal variable, as described by Webster & Oliver (2007). We then computed empirical variograms of the new transformed variables and fitted models to them. Figure 2 shows the variogram and fitted model for the transformed values of the first principal component. Note that the sill variance for the transformed variable should be 1.0. In this case the sill of the fitted model is slightly larger than this (1.04). The variogram is automatically rescaled to a sill of 1 by the disjunctive kriging program.

We then used disjunctive kriging to predict at the nodes of a square grid (interval 1 km), and for each of our seven principal components, the conditional probability that the value of the principal component falls in each of 20 intervals. These intervals were defined by the 20-percentiles of the prediction data for each principal component. The code that we used was based on that of Yates *et al.* (1986).

We then considered each of the 1000 validation data in turn. For each sample site we transformed its values for the 18 elements to the principal component scores of the PCA carried out on the prediction data set (i.e. we used the statistics of the prediction data, and the eigenvectors of their correlation matrix). We then approximated the spatial likelihood function for the sample at each node of the 1-km grid on which DK predictions were obtained. For each of the 7 principal components we identified the interval (out of 20) to which our validation sample corresponded. We then extracted the conditional probability for each of these seven intervals, and then computed the approximate spatial likelihood using Equation (**??**).

Figure 3a shows the spatial likelihood function for one of the validation sample points. It also shows the actual location of this point, note that here the true location of the point coincides with the maximum of the spatial likelihood function. However, Figure 3b shows another case where the true value did not coincide with a marked peak in the spatial likelihood function. To give an overall evaluation of the predictions by the likelihood function we proceeded as follows. First, for each validation observation we evaluated the spatial likelihood function at each of the 1-km grid nodes. We then identified the node which was closest to the actual location of the validation observation, and identified that quantile of the set of likelihoods, q_o to which the nearest node corresponded. We then computed the complement of this quantile ($c_q = 1 - q_o$), this will be zero if the nearest node is the one of maximum likelihood. We obtained c_q for each validation observation, and then plotted the empirical cumulative distribution function of this variable for the whole validation set. These numbers may be interpreted as estimates, from

the random validation sample, of the probability of including the grid node nearest to the true provenance of a sample within a subset of nodes. This subset, which constitutes proportion c_q of the full set, is designated as likely to contain the sample because the nodes have the largest values of the spatial likelihood function. The CDF of c_q is plotted, with axis labels reflecting this interpretation, in Figure 4. The solid line shows this plot for spatial likelihoods computed with 7 principal components, and the dotted line shows the effect of reducing this to the first three principal components. The dashed line is the bisector, the expected form of this plot if the spatial likelihood is only randomly related to the true location of the observation.

The plots show that the probability of including a site in a region designated from the spatial likelihood is always substantially larger than would be expected if the spatial likelihood were only randomly related to the provenance of a sample. In fact, when 7 principal components were used to determine the spatial likelihood, then if we select the top 4.5% of nodes on spatial likelihood then there is a probability of 0.5 that one of these is the nearest node to the true provenance of the sample. This proportion has to be increased to 10.9% if we only use the first three principal components. To have a probability of 0.9 that the node nearest to the true sample is included, the designated area must be 30% when we use all 7 principal components.

Discussion and Conclusions.

The case study shows that the approximate spatial likelihood has considerable potential for predicting the likely provenance of a soil sample by comparison to observations in a spatial database. It should be noted that the likelihood function may only be part of the process of inferring the provenance of soil material. Other evidence might provide us with a prior spatial probability density function. This function might, for example, exclude the possibility that the soil material came from locations further than some maximum distance from where the soil-covered exhibit was found. Integration of this prior

probability function with the likelihood function, and renormalization, would then produce a posterior probability density function which reflects how the prior distribution is rationally modified by the soil evidence expressed in the spatial likelihood function.

The case study also raises some practical issues. It is clear that there is substantial loss of information when we use just three principal components rather than 7. However, there are 18 components in total, and the plot of accumulated eigenvalues (Figure 1) shows that the variability explained by components increases more or less smoothly as the number of components is increased. It is therefore quite possible that using more than 7 principal components would give still better results. However, the process of computing Hermite transformations and modelling the variogram of many components is tedious, and is not readilly automated. Alternatively we might use indicator kriging to compute the conditional probabilities. This makes the transformation step quicker and easier; and although in theory DK retains more information, in practice little difference has been found between the estimated conditional probabilities by the two methods (Lark & Ferguson, 2004).

Two further issues require careful consideration. First, in our case study all data, both those used to estimate the spatial likelihoods (representing a database), and those for which a prediction of provenance was obtained (representing forensic specimens), had been collected on the same spatial support. The support is the particular volume, shape and orientation of the soil sample; in this case a set of five cores from the centre and vertices of a square, sampled to 15 cm depth. In practice a soil database is likely to contain data with a support similar to this, but the support of the forensic specimen is essentially unknown, since it is collected by a suspect walking over bare soil, or transferred, for example, to the wheel-arches of a vehicle from the vehicle's tyres. We distinguish, here, between the problem of unknown support and the problem of soil accumulation and mixing on an exhibit before, during and after a crime, although the latter is certainly important. Even if we could be confident that a soil specimen is from a single site the problem of

support remains. Is it soil accumulated on the specimen while walking across the site, or is it a single clod? This problem of the unknown support of soil material on forensic exhibits is of more general importance to forensic science, and Rawlins & Cave (2004) drew attention to it. In the geostatistical context it means that the probabilistic model based on soil data from a standard support will not strictly apply to the forensic data of unknown support, and since the variability of measurements will decrease as the volume of the support increases, the uncertainty attached to our statements about the forensic material is likely to be an underestimate. Further work is needed on the implications of this, and on how we might tackle the problem. It might be possible to supplement a soil database with material collected on smaller supports, to provide a variance model that can then be regularized (Journel & Huijbregts, 1978) to represent variability on a larger support as deemed appropriate.

Further, there is another potential limitation to practical implementation of the approach described in the case study. The quantity of soil retrieved from forensic items is typically very small; often less than 1 g. Significantly larger quantities of soil (*ca.* 12 g) are required for accurate, laboratory-based XRFS analyses reported in the case study. So in many real investigations it may not be possible to compare the geochemistry of the forensic sample to the database using the same analytical method. An alternative is to dissolve the sample in strong acid and analyse the resulting solution composition by ICP-MS (Inductively Coupled Plasma Mass Spectrometry); see Jarvis *et al.* (2004), but also note their reservations about ICP-MS when only small samples are available, and those of Bull *et al.* (in press). However, with the exception of the recent Tellus geochemical survey of northern Ireland (Tellus, 2007) we know of no other high-resolution soil geochemistry datasets, based on acid digest and ICP-MS analysis, which could be used as the spatial database. The general problem is whether we can reasonably compute a spatial likelihood function for a sample where the chemical analysis has been done by one method, which is different to the method used to obtain the spatial data.

To conclude, the spatial likelihood function seems to be a fruitful way of applying geostatistical inference to certain problems in forensic soil science. It provides a natural way to integrate soil with other evidence. The main problem, and one which is common to any forensic inference from soil, is how to relate the variability of reference material collected on a standard support to forensic specimens where the support is unknown and uncontrolled.

Acknowledgements

R.M. Lark's contributions were part of Rothamsted Research's programme in Mathematical and Computational Biology, which is supported by the Biotechnology and Biological Sciences Research Council of the United Kingdom. This paper is published with the permission of the Executive Director of the British Geological Survey (Natural Environment Research Council). The authors wish to acknowledge the contributions of all staff from the British Geological Survey involved in the Humber-Trent soil survey: i) the G-BASE project staff who organised the collection and processing of the soil survey data across the Humber-Trent region, ii) the staff who prepared the samples, and iii) the analytical staff who undertook the XRF analysis.

References

Bull, P.A., Morgan, R.M. & Freudiger-Bonzon, J. In Press. A critique of the present use of some geochemical techniques in geoforensic analysis (letter to the editor). *Forensic Science International*

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

CAFSS 2006. `http://www.csiro.au/resources/pfjv.html`

Finke, P.A. & Stein, A. 1994. Application of disjunctive cokriging to compare fertilizer scenarios on a field scale. *Geoderma*, **62**, 247–263.

Jarvis, K. E., Wilson, E. H. & James, S. L. 2004. Assessing element variability in small soil samples taken during forensic investigation. In *Forensic Geoscience: Principles, Techniques and Applications* (eds K. Pye & D. J. Croft), The Geological Society of London, London, pp. 171–182.

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

Lark, R.M. & Papritz, A. 2003 Fitting a linear model of coregionalization for soil properties using simulated annealing. *Geoderma*, **115** , 245–260.

Lark, R.M. & Ferguson, R.B. (2004) Mapping the conditional probability of deficiency or excess of soil phosphorous, a comparison of ordinary indicator kriging and disjunctive kriging. *Geoderma* **118**, 39–53.

Matheron, G. 1976. A simple substitute for conditional expectation: the disjunctive kriging. *In* Advanced Geostatistics in the Mining Industry (Eds. M.Guarascio, M. David and C. Huijbregts), pp 221–236. D. Reidel, Dordrecht.

McGrath, S. P. & Loveland, P. J. 1992. *The Soil Geochemical Atlas of England and Wales* Blackie Academic and Professional, Glasgow.

Pye, K. 2007. *Geological and Soil Evidence. Forensic Applications.* CRC Press, Boca Raton

Pye, K. & Blott, S. 2004. Comparison of soils and sediments using major and trace element data. In *Forensic Geoscience: Principles, Techniques and Applications* (eds K. Pye & D. J. Croft), The Geological Society of London, London, pp. 183–196.

Rawlins, B. G. & Cave, M. 2004. Investigating multi-element soil geochemical signatures and their potential for use in forensic studies. In *Forensic Geoscience: Principles, Techniques and Applications* (eds K. Pye & D. J. Croft) pp. 197-206. The Geological Society of London, London.

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

Rawlins, B. G., Kemp, S. J., Hodgkinson, E. H., Riding, J. B., Vane, C. H., Poulton, C. & Freeborough, K. 2006. Potential and pitfalls in establishing the provenance of earth-related samples in forensic investigations. *Journal of Forensic Sciences*, **51**, 832–845.

Rivoirard, J. 1994. *Introduction to disjunctive kriging and non-linear geostatistics.* Oxford University Press, Oxford.

Tellus, 2007. `http://www.bgs.ac.uk/gsni/tellus/geochemical_survey/index.html` Accessed 14th December 2007.

Webster, R. and Oliver, M.A. 2007. *Geostatistics for Environmental Scientists.* 2nd Edition. John Wiley & Sons, Chichester, UK.

Yates, S.R., Warrick, A.W. & Myers, D.E., 1986. A disjunctive kriging program for two dimensions. *Computers and Geosciences* **12**, 281–313.

Figure Captions

Figure 1 Cumulative proportion of the trace of the correlation matrix accounted for by eigenvalues of principal components of 18 elements from the Humber Trent GBASE data.

Figure 2 Empirical (symbols) variogram of the Hermite-transformed values of the first principal component of the Humber Trent data, with fitted double spherical model (line).

Figure 3 Spatial likelihood functions for two validation observations. In each case a cross indicates the actual provenance of the validation sample.

Figure 4 Estimations from the validation data of (ordinate) the probability of including the 1-km grid node closest to the true provenance of a sample in a region determined by including a specified proportion (abscissa) of sites as ordered by their spatial likelihood on (solid line) 7 or (dotted line) 3 principal components of the Humber Trent data.

Figure 1.

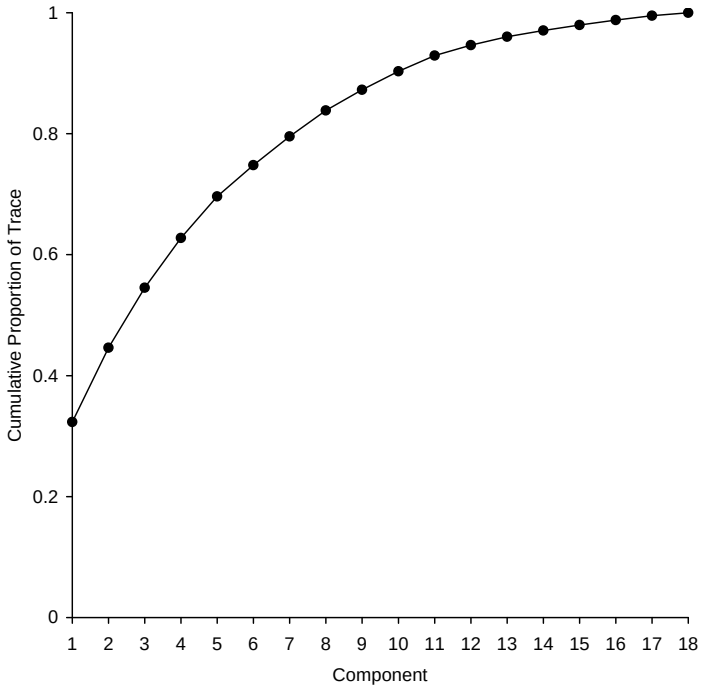

Cumulative Proportion of Trace
0
0.2
0.4
0.6
0.8
1
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Component

Figure 2.

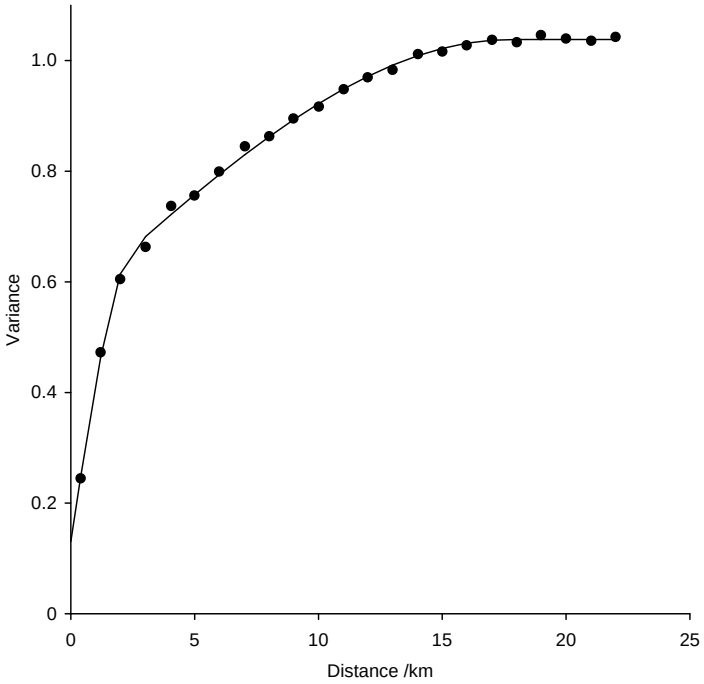


Figure 3.

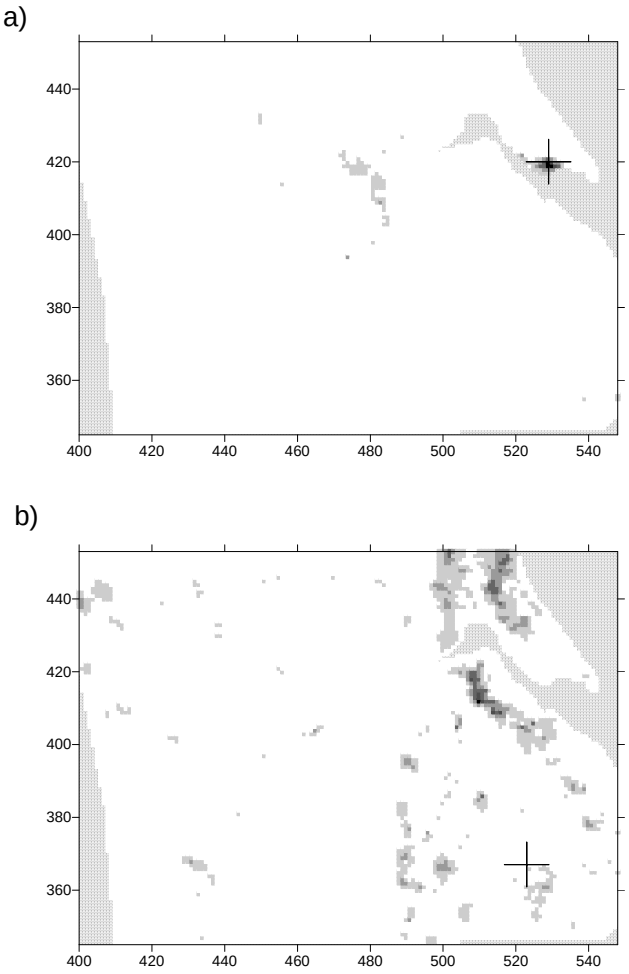
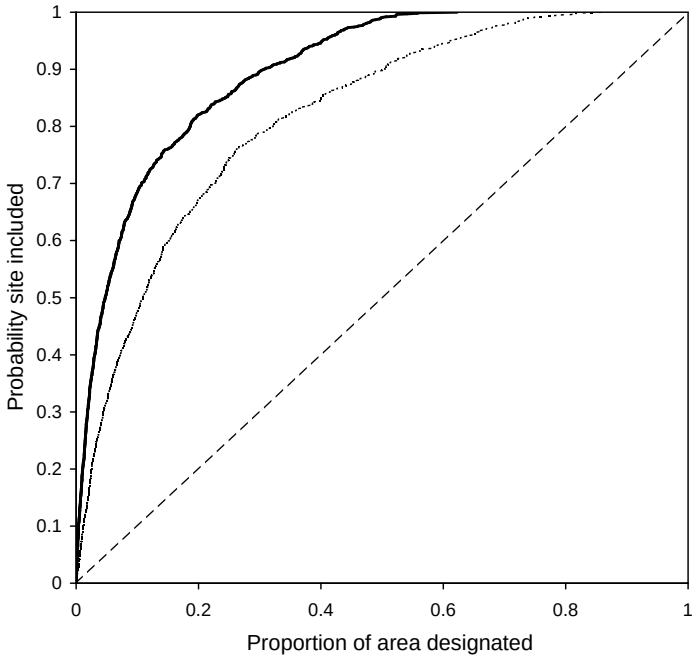

a)
440
420
400
380
360
400
420
440
460
480
500
520
540
b)
440
420
400
380
360
400
420
440
460
480
500
520
540

Figure 4.


Probability site included
Proportion of area designated
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
0
0.2
0.4
0.6
0.8
1