The assessment of point-source and diffuse soil metal pollution using robust geostatistical methods: a case study in Swansea (Wales, UK).

B. P. Marchant[a*], A. M. Tye[b] & B. G. Rawlins[b]

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

Running heading: *Point and diffuse soil metal pollution*

* Corresponding author: B.P. Marchant

e-mail: ben.marchant@bbsrc.ac.uk

Summary

The spatial variation of soil metal content arising from diffuse pollution in industrial regions cannot be analyzed by conventional geostatistical methods because predictions are influenced by metal content from natural sources and extreme values from point source pollution. We analyze a survey of soil arsenic, copper, lead, and tin at 372 locations around Swansea (Wales, UK). We use the approach of Hamon *et al.* (2004) to determine the native metal concentrations in contaminated regions from the iron content. However we find that this indicator is not appropriate around Swansea because the iron content is elevated across the contaminated region. Therefore the natural concentration of each metal is approximated by the median concentration on nearby uncontaminated rural soils on the same parent material. We divide the remaining variation between diffuse pollution and point source pollution by the robust winsorizing algorithm of Hawkins & Cressie (1984). This leads to a plausible log-Gaussian model with a constant mean which represents the diffuse pollution and estimates of the contribution of point-source pollution at each observation site. Point source pollution is found to occur at sites historically associated with production, transport and disposal of industrial wastes. The pattern of diffuse pollution is consistent with emissions from multiple smelters located throughout urban Swansea and the effects of prevailing wind and topography are evident.

Introduction

Soil contamination because of human activity has been identified as one of the major threats to soil function by the European Union in their thematic strategy for soil protection (Commission of the European Communities, 2006). National governments across the EU have separate legal frameworks for dealing with historic soil contamination. Local agencies with statutory responsibilities for the assessment and remediation of soil contamination require effective methods to map the magnitude and extent of pollution. The spatial distribution of metal and metalloid contaminants in the soil is often complex because the effects of natural sources of metals are combined with diffuse and point-source pollution. Our understanding of the processes can be enhanced by spatial predictions of the variations due to each of these three separate sources. In areas of widespread soil contamination, knowledge of the relative proportions of metal arising from natural and anthropogenic sources could aid quantitative assessments of risk to human health since the bioaccessibility of a soil contaminant can be related to the chemical form in which it entered the soil (Smith *et al.*, 2008).

Generally, regional estimates of the contribution of natural sources to metal concentrations in contaminated soil are made from the summary statistics of surveys made in areas which are assumed to be unaffected by anthropogenic processes. It is possible to distinguish between natural and anthropogenic sources of some elements such as lead by the stable isotopes (Clark *et al.*, 2006) but in other cases the metals may only have one stable isotope or analytical methods may not be widely available for the determination of isotope fractions (e.g. copper and tin). Hamon *et al.* (2004) tested whether various soil properties could be used as indicators of the background or natural metal content of contaminated soils. They found that the natural concentrations of arsenic, chromium, cobalt, copper, lead, nickel, and zinc could be approximated in terms of the iron and manganese concentrations in the soil. Their tests were conducted in south-east Asia but they suggest that these relationships may hold worldwide. This approach assumes that the iron content of contaminated soils is not elevated by anthropogenic processes. Such behaviour has been observed in previous surveys of urban soil contamination in the UK. For example, Figure 1 shows that metal processing in Sheffield has enriched the lead content of the soils in comparison with uncontaminated rural soils, but the iron content is relatively unchanged.

Conventional geostatistical methods are most efficient when the property being mapped approximates, or may be transformed to approximate, a Gaussian distribution. However point-sources of pollution lead to hotspots or outliers in the distribution of soil metals which are inconsistent with the Gaussian assumption. Therefore robust geostatistical methods have been applied to surveys of soil metal pollution. Robust methods estimate the statistics of the underlying variation of metal concentrations with minimum effect of outliers. In geostatistical analysis we first estimate a variogram model which describes the spatial variation of the property of interest based upon the observations. This model is then used to predict the property at unsampled locations. In conventional geostatistics the variogram model is estimated by Matheron's method of moments estimator (Webster & Oliver, 2007). This estimator is sensitive to outlying observations. Therefore robust variogram estimators have been devised that model the underlying variation in the presence of outliers. Three such robust estimators were compared by Lark (2000). Lark (2002) suggested a statistic which may be used to identify outlying observations. This statistic was used to identify outliers in surveys of heavy metal contamination in Sheffield, UK (Rawlins *et al.*, 2005) and Zhangjiagang, China (Zhao *et al.*, 2007). The outliers were removed from the datasets before the diffuse pollution was predicted across these study regions. However, although outliers are likely to be dominated by point-source pollution they may still contain information about the diffuse pollution. Therefore Marchant *et al.* (2010) used a robust prediction algorithm (Hawkins & Cressie, 1984) to winsorize the observations. This winsorizing process separated each observation into two components, one because of localized processes and one because of diffuse processes. A similar approach was applied by Papritz (2007) when mapping pollution around a Swiss smelter.

Although the winsorizing algorithm of Hawkins & Cressie (1984) was devised more than 25 years ago it has not been widely applied. Instead Reimann *et al.* (2005) indentified outliers in geochemical data by looking at properties of the empirical data distribution. This approach does not account for the dependence structure of the data and therefore does not explore whether the outliers are extreme relative to their nearest neighbours. The local Moran's I statistic used by Zhang *et al.* (2008) does compare each observation with its neighbours but the weight applied to each neighbour is selected arbitrarily. In contrast the winsorizing algorithm of Hawkins & Cressie (1984) ensures that the amount of influence each neighbour has is determined from a robust model of the underlying variation of the property.

In this paper we are concerned with mapping the metal content of soils around the Swansea and Neath Valleys (Wales, UK) based upon a survey of 390 observations made at 372 sites. Swansea was the world-centre of copper-smelting in the late 18$^{\text{th}}$ and early 19$^{\text{th}}$ centuries and there were other non-ferrous smelters processing arsenic, lead, zinc, silver and tin. Our aim is to quantify the effects of diffuse pollution across the study region. We test whether the natural soil content of arsenic, copper, lead and tin can be related to the concentrations of iron by conducting a second survey in a rural area that is not contaminated. We subtract our estimate of natural metal concentrations from the urban observations and separate the anthropogenic metal concentrations which remain into components due to diffuse and point-source pollution by robust geostatistical methods. This analysis yields a continuous map of diffuse metal pollution across the region and estimates of the point-source pollution at each observation site. We interpret the patterns of point-source and diffuse pollution in relation to maps of current and historical land use, and two factors which dominate deposition of airborne metals: prevailing wind and topography.

Theory

Geostatistical Prediction of Soil Properties

The variation of a soil property may be described by the linear mixed model (LMM) which divides the spatial variation between fixed and random effects (Lark & Cullis, 2004) and accounts for variation between observations made at the same site which we may think of as measurement error. The fixed effects are a linear combination of q covariates and represent variation of the expectation of the property across the study region. The random effects describe the spatially correlated component of variation of the property. The LMM is written

$$\mathbf{z} = \mathbf{M}\boldsymbol{\beta} + \mathbf{Z}\mathbf{u} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{z}$ is a length n vector of observations of the property of interest at $n_s \leq n$ distinct sites, the matrix $\mathbf{M}$ $(n \times q)$ is the design matrix for the fixed effects and contains values of the covariate at each observation site, the vector $\boldsymbol{\beta}$ of length q contains the fixed effects coefficients, the $n \times n_s$ matrix $\mathbf{Z}$ is the random effects design matrix, the vector $\mathbf{u}$ of length n_s contains the random effects and the length n vector $\boldsymbol{\varepsilon}$ contains measurement errors. The design matrix $\mathbf{Z}$ allows multiple observations from the same location to be included. If observation i is made at site j then element (i, j) of $\mathbf{Z}$ is 1. The other elements of the jth column are 0. The random effects are assumed to be a realization of a Gaussian random function U with expectation zero across the study region and covariance matrix $\mathbf{V}$. If the assumption of Gaussian underlying random effects is not plausible for a particular dataset then a transformation should be applied. The measurement errors are assumed to be independent realizations of a Gaussian function with expectation zero and variance σ_ε^2. The measurement errors can be distinguished from the nugget variation only if $n > n_s$.

The elements of $\mathbf{V}$ are obtained from a parametric function $C(\mathbf{h})$ where $\mathbf{h}$ is the lag vector separating two observations. It is common in the geostatistical literature for the spatial covariance of a random variable to be expressed in terms of the variogram

$$\gamma\left(\mathbf{h}\right) = \frac{1}{2}\mathrm{E}\left[\left\{U\left(\mathbf{x}\right) - U\left(\mathbf{x} + \mathbf{h}\right)\right\}^2\right]. \tag{2}$$

For a second order stationary random variable

$$C\left(h\right) = C\left(0\right) - \gamma\left(h\right). \tag{3}$$

The variogram may vary with both the length and direction of $\mathbf{h}$. In this paper we assume that the function is isotropic and varies only according to the length of $\mathbf{h}$ which we denote h.

A number of authorized variogram functions have been suggested which ensure that $\mathbf{V}$ is positive definite. One such example is the Matérn function (Matérn, 1960)

$$\begin{aligned} \gamma\left(h\right) &= c_0 + c_1 \left\{1 - \frac{1}{2^{\nu-1}}\Gamma\left(\nu\right)\left(\frac{h}{a}\right)^{\nu} K_{\nu}\left(\frac{h}{a}\right)\right\} \text{ for } h > 0, \\ \gamma\left(h\right) &= 0 \text{ for } h = 0, \end{aligned} \tag{4}$$

where c_0 is the nugget variance, c_1 is the partial sill variance, a is a distance parameter, ν is a smoothness parameter, K_ν is a modified Bessel function of the second kind of order ν (Abramowitz & Stegun, 1972) and Γ is the gamma function.

Conventionally the covariance parameters $\boldsymbol{\alpha} = [c_0, c_1, a, \nu, \sigma^2_\varepsilon]$ are fitted by Matheron's method of moments (Webster & Oliver, 2007). A point estimate of the variogram is made for several lag distances h based upon the mean squared difference between observations separated by lag h and a model is fitted to this point estimate by weighted least squares (Webster & Oliver, 2007). If the mean of the property varies over the study region then an initial estimate of the fixed effects coefficient can be made by least squares and the variogram is fitted to the residuals rather than the observations. Once the covariance parameters of the LMM have been fitted they may be substituted into the best linear unbiased predictor (BLUP) to calculate $\widehat{\boldsymbol{\beta}}$, an estimate of the fixed effects parameters and $\widehat{Z}(\mathbf{x_0})$ a prediction of the soil property at unobserved site $\mathbf{x}_0$. The BLUP, which is often referred to as universal kriging or kriging with external drift when fixed effects are included, also yields an estimate of the prediction variance σ^2 at each unobserved site. The BLUP predictions are weighted sums of the observations with the weights $\boldsymbol{\lambda}$ determined according to the LMM.

The validity of the fitted LMM may be confirmed by leave-one-out cross validation. For each sampling location $i = 1, \ldots n$, the value of the property at site $\mathbf{x}_i$ is predicted by the BLUP using $\mathbf{z}_{(-i)}$, the vector of observations excluding $z\left(\mathbf{x}_i\right)$ to calculate

$$\theta_i = \frac{\left\{z\left(\mathbf{x}_i\right) - \widetilde{Z}_{(-i)}\right\}^2}{\sigma^2_{(-i)}}, \tag{5}$$

where $\widetilde{Z}_{(-i)}$ and $\sigma^2_{(-i)}$ denote the prediction and prediction variance at $\mathbf{x}_i$ when $z\left(\mathbf{x}_i\right)$ is omitted from the transformed observation vector. If the fitted model is a valid representation of the spatial variation of the soil property and the prediction errors are Gaussian then $\boldsymbol{\theta} = [\theta_1 \ldots \theta_n]^{\mathrm{T}}$ is a realization of a χ^2_1 distribution with mean $\bar{\boldsymbol{\theta}} = 1.0$ and median $\breve{\boldsymbol{\theta}} = 0.455$. Quantile-quantile (QQ) plots of the $(\theta_i)^{\frac{1}{2}}$ can be drawn to confirm that the assumption of Gaussian errors is reasonable.

Robust Geostatistical Methods

The LMM representation of spatial properties assumes that the random effects can be transformed to a multivariate Gaussian distribution. However this assumption will not be plausible if the variation of a property due to an underlying process is contaminated at a small proportion of sites by a secondary process which leads to the observations at these sites being outliers. In a survey of soil metal pollution the underlying process may be the diffuse pollution and the secondary process the point-source pollution. The Matheron method of moments estimator is sensitive to outliers which lead to inflated estimates of the variance of the underlying process. Often these estimators ensure that upon cross-validation $\bar{\boldsymbol{\theta}} \approx 1.0$ but the outliers cause $\breve{\boldsymbol{\theta}}$ to be significantly less than 0.455. Outliers also have undue influence on BLUP predictions, leading to an exaggeration of the spatial extent of hotspots around an outlier.

Robust method of moments variogram estimators have been devised by Cressie & Hawkins (1980), Dowd (1984) and Genton (1998). The methods make robust point estimates of the variogram of the underlying variation. Lark (2000) tested these estimators by looking at validation statistics of variogram models fitted to simulated data. He suggested that $\breve{\boldsymbol{\theta}}$ was a suitable robust statistic to assess the fitted variograms. Lark (2000) found that Matheron's estimator outperformed the robust estimators when the property was not contaminated. However when there was contamination each of the robust estimators outperformed Matheron's estimator. The relative performance of the robust estimators varied with the form of contamination.

Lark (2002) suggested that once a robust variogram model has been fitted, outliers could be identified by a threshold on the θ_i from cross-validation. Rawlins *et al.* (2005) followed this approach and removed outliers before predicting soil metal concentrations at unsampled sites. However the removal of entire observations discards information about the underlying process. Therefore, when analysing a survey of soil metal contamination across France, Marchant *et al.* (2010) used a winsorizing algorithm suggested by Hawkins & Cressie (1984) to divide each observation between a component from underlying processes and a component from the secondary processes. They then applied the BLUP to the underlying variation and mapped the observations of the secondary process separately. The steps of this winsorizing algorithm are

1. Estimate a robust variogram of $\mathbf{z}$.

2. Compute the BLUP weights $\lambda_{j(-i)}$, $j = 1, \ldots, i-1, i+1, \ldots, n$ required for leave-one-out cross validation and the corresponding kriging variance $\sigma^2_{(-i)}$.

3. Compute the weighted median $\breve{z}_{(-i)}$ for $i = 1 \ldots n$. The weighted median solves $\sum_{j=1,j\neq i}^{n} \lambda_{j(-i)} \mathrm{sign}\left\{\breve{z}\left(\mathbf{x}_i\right) - z\left(\mathbf{x}_j\right)\right\} = 0$, where $\mathrm{sign}\left(y\right) = -1$ for $y < 0$ and $\mathrm{sign}\left(y\right) = 1$ otherwise. This equation may have more than one solution but Hawkins & Cressie (1984) state that the number of solutions is always odd and therefore a unique solution can be defined by the median of these solutions.

4. Winsorize the data by replacing z_i by
$$z_c\left(\mathbf{x}_i\right) = \begin{cases} \breve{z}_{(-i)} + c\sigma_{(-i)} & \text{if} \quad z\left(\mathbf{x}_i\right) - \breve{z}_{(-i)} > c\sigma_{(-i)} \\ z\left(\mathbf{x}_i\right) & \text{if} \quad \left|z\left(\mathbf{x}_i\right) - \breve{z}_{(-i)}\right| \leq c\sigma_{(-i)} \\ \breve{z}_{(-i)} - c\sigma_{(-i)} & \text{if} \quad z\left(\mathbf{x}_i\right) - \breve{z}_{(-i)} < -c\sigma_{(-i)} \end{cases} \tag{6}$$
where c is a constant $1.5 < c < 3.0$.

5. Predict the property at unsampled locations by application of the BLUP to $\mathbf{z}_c$ rather than $\mathbf{z}$.

Marchant *et al.* (2010) repeated the above algorithm for different values of c and calculated cross-validation $\boldsymbol{\theta}$ statistics for each $\mathbf{z}_c$. The use of a robust variogram estimator in stage 1 ensured that for large c, $\breve{\boldsymbol{\theta}} \approx 0.455$ but in the presence of outliers $\bar{\boldsymbol{\theta}} > 1.0$. The value of $\bar{\boldsymbol{\theta}}$ decreased more rapidly than $\breve{\boldsymbol{\theta}}$ as c was decreased and their final prediction of the underlying variation was based upon the $\mathbf{z}_c$ for which $\bar{\boldsymbol{\theta}}$ was closest to 1.0. In the original formulation of the Hawkins & Cressie (1984) algorithm the mean of $\mathbf{z}$ was assumed to be constant and the BLUP in Step 2 was equivalent to ordinary kriging. Papritz (2007) expanded the algorithm to include fixed effects. The fixed effect coefficients were estimated by a robust regression estimator and the winsorizing algorithm was applied to the residuals.

Methods

The Study Area

The study region encompasses an area of south Wales (UK) shown in Figure 2 with the underlying soil parent materials (British Geological Survey, 2006). Figure 3 shows the urban area of Swansea and includes the topographic features such as the Swansea and Neath Valleys which extend to the north and north-east from Swansea Bay. For the wider study region, where bedrock is the parent material, it is dominated by medium to coarse-grained sandstone of the Penant Sandstone Formation, which also comprises claystones, siltstones and minor fine-grained sandstones that contain coal seams. The glacial tills are mostly associated with the Late Devensian glaciation including clasts of Old Red Sandstone and Carboniferous Limestone from the Brecon Beacons. In the Swansea Valley, the till deposits are overlain by glaciolacustrine deposits which include clay and silt (Figure 3). Glaciolacustrine deposits also occupy the Neath Valley, including sand and gravel deposits. During the Holocene, alluvium was deposited and peat deposits formed in upland and lowland areas of restricted drainage. The dominant soils across the study region have been described as fine loamy soils, sometimes with slight waterlogging (Soil Survey of England and Wales, 1983).

In late 18^{th} and early 19^{th} century Swansea there were many smelters processing copper, arsenic, lead, zinc, silver and tin. The height of the chimney stacks was increased in the 19^{th} century to disperse the toxic fumes from the copper smelters.

The lead-smelting industry was particularly significant in the 17th and 19th centuries, although compared to copper a greater proportion of smelting was undertaken in the ore fields. A total of 250 000 tonnes of raw copper-ore was processed in the Swansea Valley annually in the mid 19th century yielding 22 000 tonnes of refined copper; the dominant source of ore was Devon and Cornwall (Hughes, 2000). The copper industry was considered to be the principal contributor to Swansea's pollution problems. Newell & Watts (1996) used a Gaussian plume model to estimate annual average suspended airborne concentrations of arsenic, lead and tin during the mid-19th century in the vicinity of the Llanelli copper company 12 miles north-west of Swansea. The estimates were between 10 and 15 $\mu g\ m^{-3}$. By contrast, current EC regulations stipulate limits of 2 $\mu g\ m^{-3}$. More recently remediation has been undertaken; the Lower Swansea Valley project of the 1960s and 1970s reclaimed slag heaps and large tracts of derelict land.

The Urban Survey

Soil samples were collected in 1994 from 372 sites around Swansea on a regular grid at a density of four sites per square kilometre (Figure 3). Marchant & Lark (2007a) and Marchant & Lark (2007b) showed that the efficiency of regular grid surveys could be greatly improved if a few additional samples were collected from sites close to sites on the regular grid. These additional samples lead to a more accurate estimate of the variogram over small lag distances. Therefore additional samples were collected 20 m away from six of the regular grid sites. At these six sites both the sample from the grid site and the additional sample 20 m away were split into two subsamples to allow measurement errors to be explored. Thus a total of 390 samples were collected.

Samples were collected according to the protocols of the Geochemical Surveys of Urban Environments (GSUE) project (Fordyce *et al.*, 2005) across Swansea, Neath, Port Talbot and the Mumbles area of the Gower Peninsula. Sample sites were selected from open ground as close as possible to the centre of each of four 500-metre squares, within each kilometre square of the British National Grid (BNG). Typical locations for sampling were gardens, parks, sports fields, road verges, allotments, open spaces, schoolyards and waste ground. Each composite sample was based on nine samples of equal size from the corners, sides and centre of square of side-length 2 m. Each sample was collected at a depth range of 0-15 cm from the soil surface using an auger of diameter 35 mm. At each site, information was recorded on location using 1:10 000 scale Ordnance Survey maps, a description of any visible contamination (e.g. metallic, pottery, bricks, plastics etc.), Munsell colour, soil clast lithologies (e.g. sandstone, limestone, etc.) and land use. All soil samples were disaggregated following air-drying and sieved to less than 2 mm. All samples were coned and quartered, and a 50-g subsample was ground in an agate planetary ball mill. The total concentrations of 18 major and trace elements were determined by wavelength and energy dispersive X-ray fluorescence spectrometry (XRF-S). In this paper we only consider five elements (detection limits in parentheses): arsenic (1 mg kg^{-1}), copper (1 mg kg^{-1}), total iron expressed as Fe_2O_3 (0.01 %), lead (2 mg kg^{-1}), and tin (1 mg kg^{-1}). For brevity we refer to these variables as metal concentrations although arsenic is a metalloid. Brief descriptions of the local land use at and around each site were tabulated for the years 1900 and 2007 from Ordnance Survey maps of the area.

The Rural survey

The sampling locations for the rural survey are shown in Figure 2. In selecting the area in which to locate sampling sites we wished (i) to avoid the effects of atmospheric metal deposition in the vicinity of Swansea, giving consideration to the prevailing south and south-westerly wind directions (ii) to avoid the influence of other smaller urban areas around Swansea and (iii) to ensure the soils were derived from the same dominant parent material types that are found around Swansea (the Penant Sandstone Formation and glacial till).

We selected an area approximately 25 km to the west of Swansea where these conditions were met; this area is also 2 km downwind of the coast, ensuring minimal atmospheric sources of metal. We chose to sample the soil at 23 sites; 15 sites over sandstone parent material and eight sites over areas where glacial till had been mapped

(British Geological Survey, 2006). The precise sampling locations were randomly selected although limitations in access to sites due to crops and livestock were taken into account. The soil samples were collected in January 2007. At each sampling site, five incremental soil samples were collected using a Dutch auger at the corners and centre of a square with a side of length 20 m and combined to form a composite sample of approximately 0.5 kg. At each of these five points, any surface litter was removed and the soil sampled to a depth of 15 cm into the exposed soil. On return to the laboratory, the same preparation and analytical protocols were applied to each sample as those described above for the urban survey.

Statistical Analysis of Soil Metal Concentrations Around Swansea

We assume that the spatial variation of soil metal concentrations in the urban soil is the sum of three factors, (i) natural sources of metals (ii) diffuse pollution (iii) point-source pollution. We attempted to separate these three components of variation. The variation due to natural sources was modelled from the rural observations. Regression analyses were conducted on the rural observations to evaluate the relationships between the four metals of interest and the total iron concentration as suggested by Hamon *et al.* (2004). Also, the empirical cumulative distribution function (CDF) for the rural iron observations was compared with the corresponding CDF from the Swansea urban survey to determine whether the soil iron concentration has been enriched in Swansea.

The predicted contribution of natural sources to the observed soil metal concentrations was subtracted from the total urban observation to leave the observed component due to anthropogenic processes. These anthropogenic observations were highly skewed and therefore the data were log-transformed. The components due to diffuse pollution and point-source pollution were separated by robust geostatistical methods. The approach was broadly similar to that applied by Marchant *et al.* (2010) when mapping metals across France. Matérn variograms were fitted to the anthropogenic observations of each metal by the method of moments in conjunction with Matheron's estimator and the robust estimators suggested by Cressie & Hawkins (1980), Dowd

(1984) and Genton (1998). Cross-validation was performed for each fitted variogram and the estimator with $\breve{\boldsymbol{\theta}}$ closest to 0.455 was selected. The observations were then winsorized according to the algorithm of Hawkins & Cressie (1984) for various values of constant $1.5 < c < 3.0$. This algorithm removes both positive and negative outliers. However we expect that the majority of outliers will be positive and caused by point source pollution. Therefore we only remove these positive outliers.

The mean of $\boldsymbol{\theta}$ was calculated for each c and the winsorized observations $\mathbf{z}_c$ for which $\bar{\boldsymbol{\theta}}$ was closest to 1.0 were assumed to be observations of the diffuse pollution. The $\mathbf{z}_c$ observations were predicted across the study region by the BLUP with a global search neighbourhood and these predictions were back-transformed to the original units by the exponential transform. We note that this leads to an estimate of the median rather than the mean in the original units. We consider the median to be the more appropriate statistic for a contaminated dataset. The difference between the anthropogenic observations and the observations of the diffuse pollution were assumed to be the effect of point-source pollution.

We note that the choice of robust variogram estimator was based upon non-robust cross validation statistics. The $\breve{\boldsymbol{\theta}}$ statistic could have been assessed after the observations had been winsorized but this would lead to an excessive number of computations since it would require that the winsorizing algorithm was applied for each of the four robust variograms and a range of c values.

Results

Prediction of Natural Metal Concentrations

Table 1 shows the summary statistics of the rural soil metal concentrations and the correlations between these metals and total iron. In each case these correlations are small and the p-values for the null hypothesis that the metal concentrations are independent of the total iron content are greater than 0.4. Additionally, the empirical CDFs (Figure 4) demonstrate that iron concentrations are greater throughout the urban survey than in the rural survey. Both of these findings indicate that the method of

Hamon *et al.* (2004) for determination of the component of the metal concentrations due to natural sources is not appropriate for this study. Therefore we approximate the natural concentration of each metal by its median in the rural survey (Table 1).

Geostatistical Prediction of Anthropogenic Metal Concentrations

The Matheron and robust variograms fitted to each log-transformed metal are compared in Figure 5. For the anthropogenic component of each of the metals the cross-validation statistics for the Matheron variogram had $\breve{\boldsymbol{\theta}} < 0.455$ (Table 2) and therefore the variogram was not valid. In each case $\breve{\boldsymbol{\theta}}$ increased to a value closer to 0.455 when a robust estimator was used. The $\bar{\boldsymbol{\theta}}$ value was greater than 1.0 for each of the robust estimators. However it was possible to select a winsorizing constant $1.5 < c < 3.0$ such that $\bar{\boldsymbol{\theta}}$ for the winsorized component $\mathbf{z}_c$ was approximately 1.0. The values of $\breve{\boldsymbol{\theta}}$ for the winsorized component were in the range $0.4 \leq \breve{\boldsymbol{\theta}} \leq 0.455$. Our use of the $\breve{\boldsymbol{\theta}}$ statistic to assess the suitability of the models assumes that the prediction errors are Gaussian. We confirm that this assumption is reasonable with QQ plots (Figure 6). For the robust variogram fitted to the uncensored observations the majority of standardized errors lie close to the $x = y$ line and indicate that it is reasonable to assume that the prediction errors for the underlying variation are Gaussian. A number of prediction errors deviate from the $x = y$ line at both extremes of the distribution. However by censoring only the positive outliers all these errors move closer to the $x = y$ line. This indicates that the negative outliers are artefacts. They are located close to positive outliers and are only outliers relative to these observations. After winsoring all of the prediction errors for copper and arsenic are close to the $x = y$ line. For lead and tin it appears that the winsorizing process has removed too much of the observation. The predicted maps of the metal concentrations because of diffuse pollution (the censored observations) and the observations of the point-source pollution (the difference between the observations and the censored observations) are shown in Figure 7.

Distribution and magnitude of point and diffuse metal pollution

There are some common features in the maps of diffuse pollution of each metal. In each, the long-axis of the areas with elevated concentrations is consistent with the prevailing wind direction (oriented approximately 225° clockwise from north). Diffuse pollution is elevated on the western side of the Swansea Valley and within the wider Neath Valley. Less pollution is evident on the western edge of the study region. The lead and tin diffuse pollution is concentrated into a few localized regions whereas larger areas of elevated copper and arsenic diffuse pollution are evident. The pattern of arsenic diffuse pollution is dominated by one large area to the south-east of the Swansea Valley.

Of the four metals, copper has the most sites at which point-source pollution is evident. Local details from Ordnance Survey maps of recent (2007) and historic (1900) land use at the sites affected by point-source pollution are presented in Table 3. Land use at or around the vast majority of these sites is associated with either production (works), transport (railways and docks) or potential disposal (collieries and quarries) of industrial wastes. At two sites where large concentrations of lead were reported (2768 and 3942 mg kg $^{-1}$) the land use information does not indicate any local source for the metal; the latter site was recorded as a domestic garden during the survey which could be of some concern given the potential implications for human health through exposure to lead in the soil.

Discussion

The survey confirms that the soils around Swansea remain substantially contaminated by historic metal and metalloid pollution. The soil metal concentrations cannot be represented by conventional geostatistical methods because the combination of diffuse and point-source pollution leads to complex patterns of variation. When conventional models were fitted to the data they were found to be invalid. The estimated variances were inflated by a small number of large observations at former industrial sites and thus it was not possible to accurately quantify the uncertainty of the predictions which result. However, plausible models did result when the diffuse and point-source pollution were mapped separately by robust geostatistical methods. In a previous survey, robust methods were also required to map diffuse metal pollution around Sheffield (Rawlins *et al.*, 2005) and it is likely that that similar methods will be required to assess metal contamination in other industrial regions.

It was not possible to map the variation of the natural metal content of the soil. A relationship between natural metal concentrations and total iron in the soil suggested by Hamon *et al.* (2004) does not apply in this study region. However since the variation of metals from natural sources in this survey was dwarfed by the anthropogenic contribution it was adequate to assume that the natural concentration of each metal was constant across the study region and approximate it by the median concentration in a nearby uncontaminated rural area.

Documentary evidence suggests that the majority of the diffuse metal pollution across Swansea was the result of atmospheric deposition of metals to the soil following their dispersal from smelter stacks (Hughes, 2000). The patterns of diffuse pollution are consistent with emissions from numerous smelters located throughout the urban areas. The patterns are influenced by the topography of the region and the prevailing wind direction. The spatial predictions could potentially be improved if these factors are included in a process model of deposition following atmospheric dispersal from specific sources across the region.

The model used in this study assumed a constant mean across the study region. Once the winsorizing had been completed a LMM including fixed effects could have been fitted to the censored observations. We did test models where elevation and parent material were included as fixed effects. However modified likelihood tests (Marchant *et al.*, 2009) suggested that these did not lead to a significantly improved fit. We suggest that elevation is not a suitable fixed effect because the amount of contamination differs on each side of the valleys and that the proximity of a source of contamination is a more important factor than the parent material. Anisotropy could also have been added to the model at this stage.

The pattern of sites where point-source pollution was identified is consistent with metal production, transport and disposal occurring at numerous sites across the urban area. We note that the robust algorithm identifies local outliers as well as global outliers. Local outliers are not necessarily extreme in comparison with the whole dataset but are extreme in comparison to neighbouring observations. For example one copper observation has been identified as an outlier despite the concentration only being 100 mg kg^{-1}. This is because there was a second observation from the same site of 40 mg kg^{-1}. Such outliers would not be found by algorithms based upon the empirical data distribution (Reimann *et al.*, 2005).

There were some differences between the soil contamination observed in Swansea and that previously observed in Sheffield (Rawlins *et al.* 2005). Elevated concentrations of total iron were observed throughout urban Swansea but not urban Sheffield. We hypothesise that the difference between the situations in Swansea and Sheffield are because Sheffield was a centre of metal processing whereas Swansea was a centre of metal smelting. Therefore more ferrous waste was brought into Swansea within the metal ores. Also, the median concentration of lead in topsoil from diffuse pollution in the survey of Swansea (180 mg kg^{-1}) was substantially larger than the value of 73 mg kg^{-1} (urban median of 161 mg kg^{-1} minus rural median of 88 mg kg^{-1}) reported by Rawlins *et al.* (2005) in Sheffield. These estimates are comparable because in each case statistical outliers or hotspots in the urban area were removed from the data. We believe that the substantially larger concentrations of lead across Swansea – in comparison to Sheffield – result from atmospherically deposited metal due to smelting of metal ores within the urban area of Swansea.

In England and Wales the first tier of a human health or ecological risk assessment is a comparison between observed total soil metal concentrations at a site and their guideline values (Environment Agency, 2009) or screening values (Environment Agency, 2008). In the case of human health risk assessment, the revised Soil Guideline Values for arsenic concentrations in topsoil (32 mg kg $^{-1}$ for residential land use) is exceeded by the predicted sum of natural content and diffuse pollution for 89% of the study area. Ecological health risks are assessed according to the difference between observed concentrations and ambient background metal concentrations (ABC) in soil. The proposed screening values for lead (167 mg kg $^{-1}$) and copper (88 mg kg $^{-1}$) are exceeded by the predictions of diffuse pollution for 44% and 58% of the study area respectively. When the ABCs are established it is important to ensure that they do not include any diffuse metal pollution.

Exposure to soil Pb can also occur through inhalation of airborne particulates. Average monthly Pb concentrations (ng m^{-3}) of fine (PM_{10}), particulates measured during 2008 in air from sites in Swansea (Swansea Coedgwilym – 8) and another in Port Talbot (Port Talbot Margam – 11.9) were below the average of 16 ng m^{-3} from all 24 sites in the UK Heavy Metals Monitoring Network (see Brown *et al.*, 2010). Another site in Swansea (Morriston) had annual average concentrations of particulate Pb in air of 20.5 ng m^{-3}, somewhat greater than the national average. Although there is some evidence that the enhanced concentrations of topsoil Pb concentrations across Swansea may enhance its concentration in airborne particulates, the overall relationship is complex and requires further study.

Conclusions

This study illustrates that when soil properties are mapped it is vital to validate the statistical model of the property to ensure that it is appropriate. Conventional geostatistical models were not appropriate for the prediction of diffuse soil metal contamination across urban Swansea because the estimated variograms and predictions were overly influenced by point source pollution. However these different components of contamination were separated and mapped by robust geostatistical methods. The large concentrations of tin, lead, copper and arsenic in topsoil across the urban Swansea area have significant implications for human health and ecological risk assessments according to current guidance for England and Wales. The methods described in this paper are likely to be required to map soil pollution around other industrial centres.

Acknowledgements

This paper is published with the permission of the Executive Director of the British Geological Survey (Natural Environment Research Council). We acknowledge the contributions of all staff from the British Geological Survey involved in the soil geochemical survey of Swansea and the XRF-S analysis. BPM's contribution is part of Rothamsted Research's program in Mathematical and Computational Biology funded by the Biotechnology and Biological Sciences Research Council through its strategic grant to Rothamsted Research.

References

Abramowitz, M. & Stegun, I.E. (Eds) 1972. *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* 10th Printing. U.S. Department of Commerce, National Bureau of Standards, Washington DC.

British Geological Survey 2006. *Digital Geological Map of Great Britain 1:50 000 scale (DiGMapGB-50) data [CD-ROM] Version 3.14.* British Geological Survey, Keyworth, Nottingham.

Brown, R.J.C. 2010. Comparison of estimated annual emissions and measured annual ambient concentrations of metals in the UK 1980–2007. *Journal of Environmental Monitoring,* **12**, 665-671.

Clark, H.F., Brabander, D.J. & Erdil, R.M. 2006. Sources, sinks, and exposure pathways of lead in urban garden soil. *Journal of Environmental Quality*, **35**, 2066–2074.

Commission of the European Communities, 2006. Thematic Strategy for Soil Protection. Brussels. http://ec.europa.eu/environment/soil/pdf/com_2006_0231_en.pdf. Accessed 25th March 2009.

Cressie, N. & Hawkins, D. 1980. Robust estimation of the variogram. *Mathematical Geology*, **12**, 115–125.

Dowd, P.A. 1984. The variogram and kriging: robust and resistant estimators. In: *Geostatistics for Natural Resources Characterization* (eds Verly, G., David, M., Journal, A.G. & Marechal, A.). Part 1, p. 91–106. D. Reidel, Dordrecht.

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

Environment Agency, 2009. Soil screening values for assessing ecological risks. Accessed 17th July, 2009. http://www.environment-agency.gov.uk/static/documents/Research/ssv_2149429.pdf

Fordyce, F.M., Brown, S.E., Ander, E.L., Rawlins, B.G., O'Donnell, K.E., Lister, T.R. Breward, N. & Johnson, C.C. 2005. GSUE: urban geochemical mapping in Great Britain. *Geochemistry: Exploration, Environment, Analysis*, **5**, 325–336.

Genton, M.G. 1998. Highly robust variogram estimation. *Mathematical Geology*, **30**, 213–221.

Hamon, R.E., McLaughlin, M.J., Gilkes, R.J., Rate, A.W. Zarcinas, B., Robertson, A., Cozens, G., Radford, N. & Bettenay, L. 2004. Geochemical indices allow estimation of heavy metal background concentrations in soils. *Global biogeochemical cycles* **18**, GB 1014.

Hawkins, D. M. & Cressie, N. 1984. Robust kriging — A proposal. *Mathematical Geology*, **16**, 3–18.

Hughes, S. 2000. *Copperopolis – landscapes of the early industrial period in Swansea.* Royal Commission on the ancient and historical monuments of Wales. Cambrian Printers Limited, Ceredigion. pp. 358.

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

Lark, R.M. 2002. Modelling complex soil properties as contaminated regionalized variables. *Geoderma*, **106**, 173–190.

Lark, R.M. & Cullis, B.R. 2004. Model based analysis using REML for inference from systematically sampled data on soil. *European Journal of Soil Science*, **55**, 799–813.

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

Marchant, B.P. & Lark, R.M. 2007b. Optimal sampling for geostatistical surveys. *Mathematical Geology*, **39**, 113–134.

Marchant, B.P., Newman, S., Corstanje, R., Reddy, K.R., Osborne, T.Z. & Lark, R.M. 2009. Spatial monitoring of a non-stationary soil property: phosphorus in a Florida water conservation area. *European Journal of Soil Science*, **60**, 757–769.

Marchant, B.P., Saby, N.P.A., Lark, R.M., Bellamy, P.H., Jolivet, C.C. & Arrouays, D. 2010. Robust prediction of soil properties at the national scale: Cadmium content of French soils. *European Journal of Soil Science*, **61**, 144–152.

Matérn, B. 1960. Spatial variation. Meddelanden från Statens Skogsforskningsinstitut, 49, No. 5. [2nd Edition (1986), Lecture Notes in Statistics, No. 36, Springer, New York].

Newell, E. & Watts, S. 1996. The environmental impact of industrialization in South Wales in the Nineteenth century:'Copper smoke' and the Llanelli Copper Company. *Environment and History*, **2**, 309–336.

Papritz, A. 2007. Robust universal kriging. *Pedometrics 2007*, Tuebingen, Germany p. 15.

Rawlins, B.G., Lark, R.M., O'Donnell, K.E., Tye, A.M. & Lister, T.R. 2005. The assessment of point and diffuse metal pollution from an urban geochemical survey of Sheffield, England. *Soil Use and Management*, **21**, 353–362.

Reimann, C., Filzmoser, P. & Garrett, R.G. 2005. Background and threshold: critical comparison of methods of determination. *Science of the Total Environment*, **346**, 1–16.

Smith, E., Naidu, R., Weber, J. & Juhasz, A. L. 2008. The impact of sequestration on the bioaccessibility of arsenic in long-term contaminated soils. *Chemosphere*, **71**, 773–780.

Soil Survey of England and Wales 1983. *Soils of Wales.* Ordnance Survey for the Soil Survey of England & Wales, Southampton.

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

Zhao, Y.C., Xu, X.H., Huang, B., Sun, W.X., Shao, X.X., Shi, X.Z. & Ruan, X.L. 2007. Using robust kriging and sequential Gaussian simulation to delineate the copper- and lead-contaminated areas of a rapidly industrialized city in Yangtze River Delta, China. *Environmental Geology*, **52**, 1423–1433.

Zhang, C., Luo, L., Xu, W. & Ledwith, V. 2008. Use of local Moran's I and GIS to identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of the Total Environment*, **398**, 212–221.

Figure captions

Figure 1: Empirical cumulative density functions of metal concentrations in urban soil of Sheffield (n=588 sites) and soil of surrounding rural areas (n=818 sites) developed over the same parent material type (Coal Measures): a) iron and b) lead (Pb). For further details see Rawlins *et al.* (2005).

Figure 2: Parent materials across the study region in relation to Swansea (shown in outline) and the soil sampling locations for estimation of natural metal concentrations (n=23).

Figure 3: Soil sampling locations (n=373) in Swansea and their parent materials types superimposed on a digital elevation model. Grid coordinates are metres of the British National Grid.

Figure 4: Empirical cumulative density functions of iron concentrations in urban soil of Swansea (n=373 sites; sampled in 1994) and rural sites (n=23 sites; sampled in 2007).

Figure 5: Matheron (dashed curves and '.'s) and best robust variograms (continuous curves and 'x's) for log-transformed metal concentrations.

Figure 6: QQ plots for the standardized prediction errors from a robust variogram for the transformed observations (left) and the winsorized transformed observations (right).

Figure 7: Predicted maps of diffuse metal pollution (a), (c), (e) and (g) and point-source metal concentration (b), (d), (f) and (h). Labels on locations of point-source pollution correspond to entries of Table 3. The origin of the maps is a British national grid reference 260000, 187000 and the ticks denote 5000-m increments.

Table 1 Summary statistics of metal concentrations at sites for usual background value sites (UBV; n=23) and from the urban survey of Swansea (USS; n=373). Units mg kg^{-1} unless stated.

Element	As		Cu		Fe_2O_3 (%)		Pb		Sn	
Dataset	UBV	USS	UBV	USS	UBV	USS	UBV	USS	UBV	USS
Mean	31.3	76.8	36.1	161	3.99	6.29	49.63	432	7.6	58
Median	30.2	53.0	35.7	114	3.97	5.92	48.0	224	7.3	31
Standard deviation	15.0	126.7	11.1	173	0.90	2.34	13.9	926	2.61	92
Skew	2.89	11.0	1.09	4.01	0.31	1.89	1.01	11.0	2.07	5.39
Correlation with Fe	0.1		0.09		1		-0.06		-0.18	
[a]P-value	0.67		0.65		0		0.78		0.41	

[a] P-value for null hypothesis that variable is independent of Fe_2O_3.

Table 2 Cross-validation statistics for variograms fitted by Matheron's estimator and the best robust estimator.

	Cu	As	Pb	Sn
[a] $\bar{\boldsymbol{\theta}}_{\mathrm{M}}$	1.15	1.03	0.88	0.97
$\breve{\boldsymbol{\theta}}_{\mathrm{M}}$	0.35	0.39	0.30	0.40
Estimator	Dowd	Genton	Dowd	Dowd
[b] $\bar{\boldsymbol{\theta}}_{\mathrm{R}}$	1.40	1.19	1.03	1.15
$\breve{\boldsymbol{\theta}}_{\mathrm{R}}$	0.44	0.46	0.41	0.44
c	2.1	2.3	2.7	2.4
[c] $\bar{\boldsymbol{\theta}}_{c}$	1.01	1.01	1.00	1.00
$\breve{\boldsymbol{\theta}}_{c}$	0.40	0.44	0.41	0.44

[a] $\boldsymbol{\theta}_{\mathrm{M}}$ cross-validation statistic for Matheron estimator

[b] $\boldsymbol{\theta}_{\mathrm{R}}$ cross-validation statistic for best robust estimator

[c] $\boldsymbol{\theta}_{c}$ cross-validation statistic for winsorized data

Table 3 Land use (current and historic) types for point-source metal and metalloid contaminants (soil concentration in mg kg^{-1}). References correspond to labels in Figure 6. Features next to land use (derived from Ordnance Survey maps) shown in parentheses.

Ref.	Concentration	Land use at given date	
		2007	1900
Cu			
1	323	Grassland	No detail on map
2	1160	Waste ground (railway)	Field close to steelworks and colliery
3	100	Field	Field close to colliery
4	1119	Domestic garden	Railway Yard
5	354	Waste ground (railway)	Close to railway; Close to Morriston spelter works; Railway yard
6	999	Waste ground (railway)	Railway Yard and Swansea Chemical works
7	1477	Path (river, quarries, works)	Close to Ni and Co works; close to station
8	1297	Railway	Close to canal tow path and railway yard
9	1149	Docks	Below high water mark
10	667	Docks/Landing stage (works)	Baglam Bay - No development, next to river Neath
11	259	Industrial estate	Field, adjacent to railway
12	172	Ground around housing	Ground around housing
13	376	Ground around housing	Ground around housing
As			
14	407	Field (quarry)	Field close to pit
7	917	Path (river, quarries, works)	Close to railway; Close to Morriston spelter works; Railway yard
15	2047	Quarry	Field
9	398	Docks	Below high water mark
16	214	Close to railways	Railway sidings
17	501	Field adjacent to colliery	Field (grassland)
Pb			
18	3942	Domestic garden	Domestic Garden
19	2768	Domestic garden	Field
9	6075	Docks	Below high water mark
Sn			
2	351	Waste ground (railway)	Field close to steelworks and colliery
20	553	Industrial estate	Tin plate works
21	919	Field (pit)	Field close to brick works and quarry
7	834	Path (river, quarry, works)	Close to railway; Close to Morriston spelter works; Railway yard
22	329	Quarry	Field
9	452	Docks	Below high water mark
23	99	Railway	Industrial estate

Figure 1:

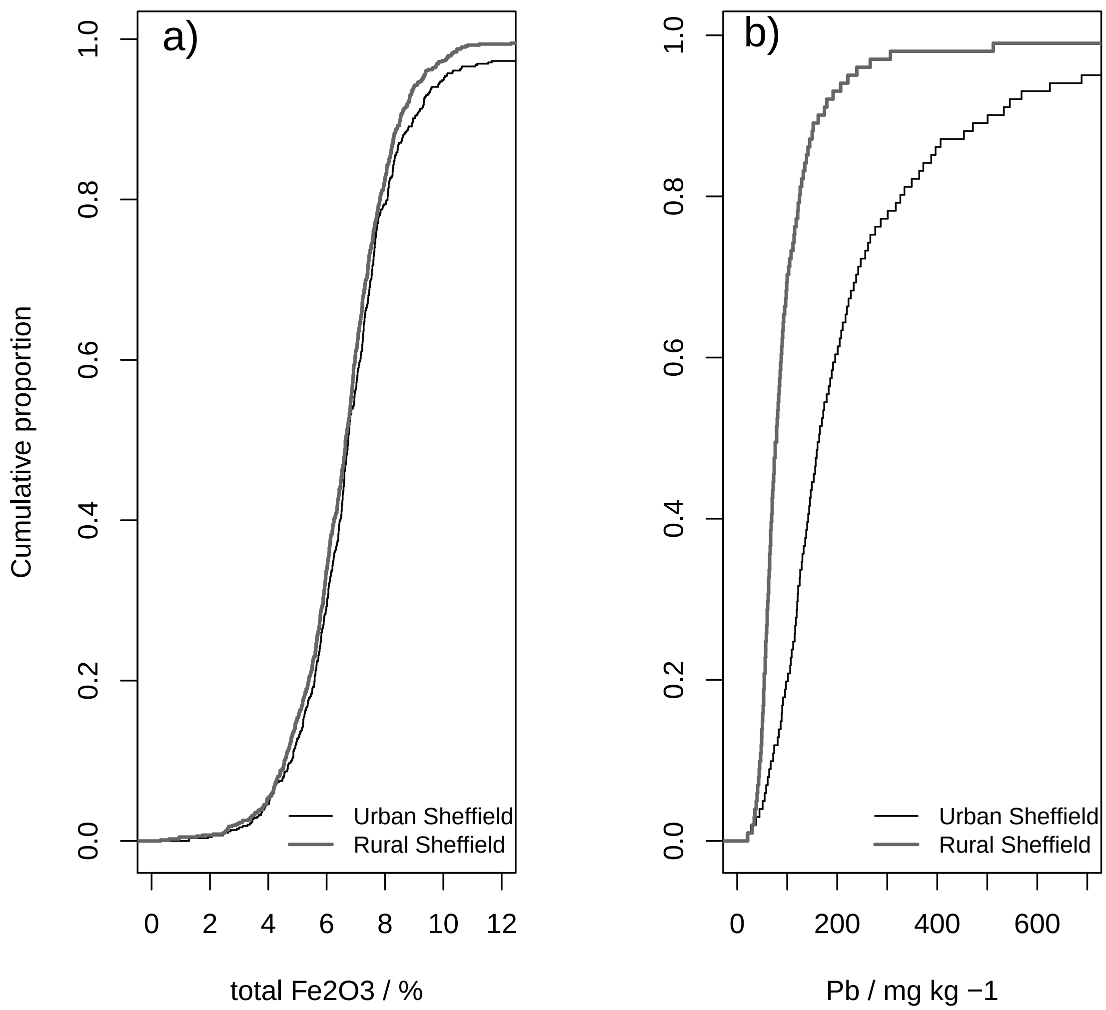

a)
b)
Cumulative proportion
0.0
0.2
0.4
0.6
0.8
1.0
0
2
4
6
8
10
12
200
400
600
total Fe2O3 / %
Pb / mg kg −1
Urban Sheffield
Rural Sheffield

Figure 2:

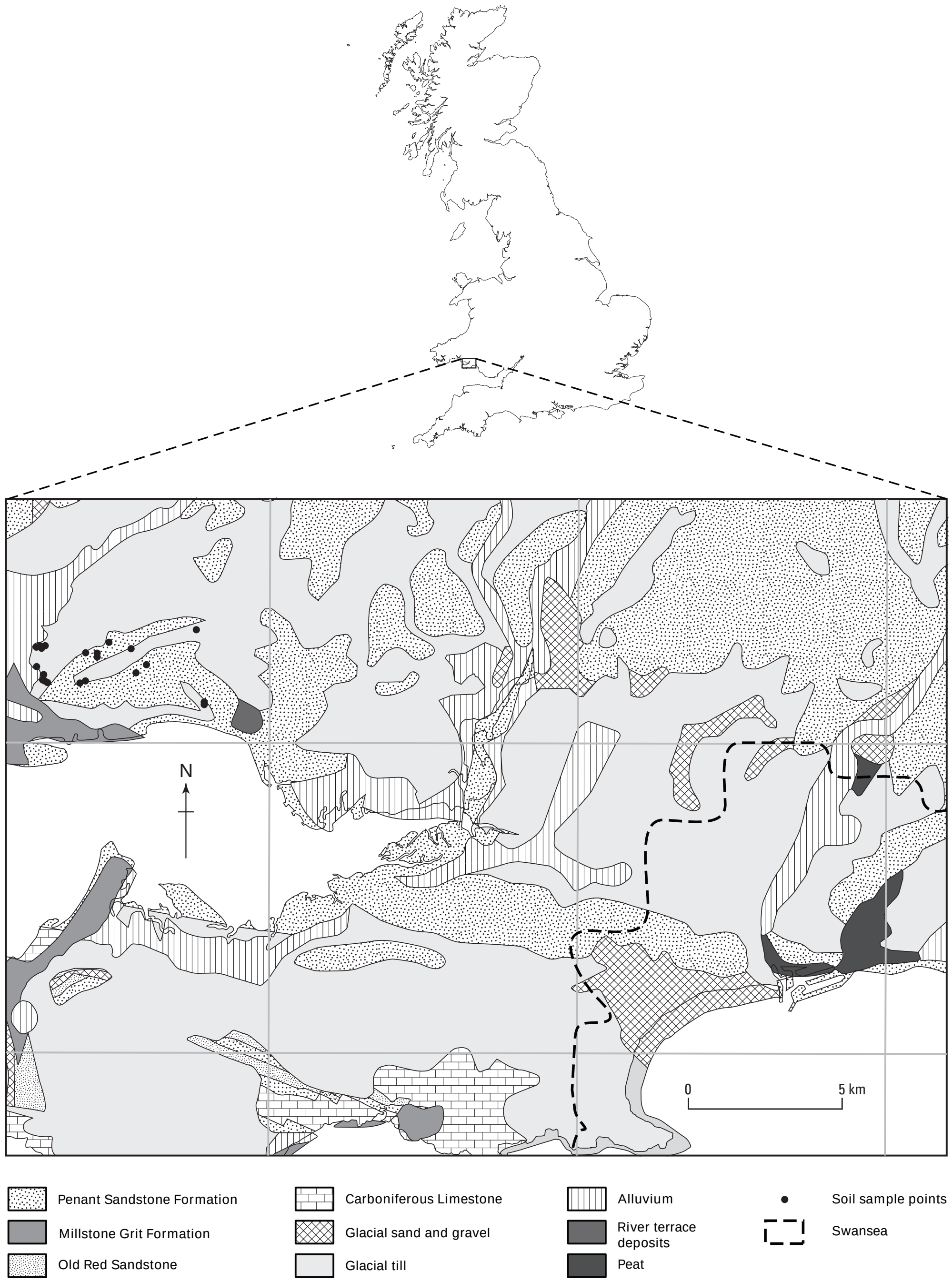
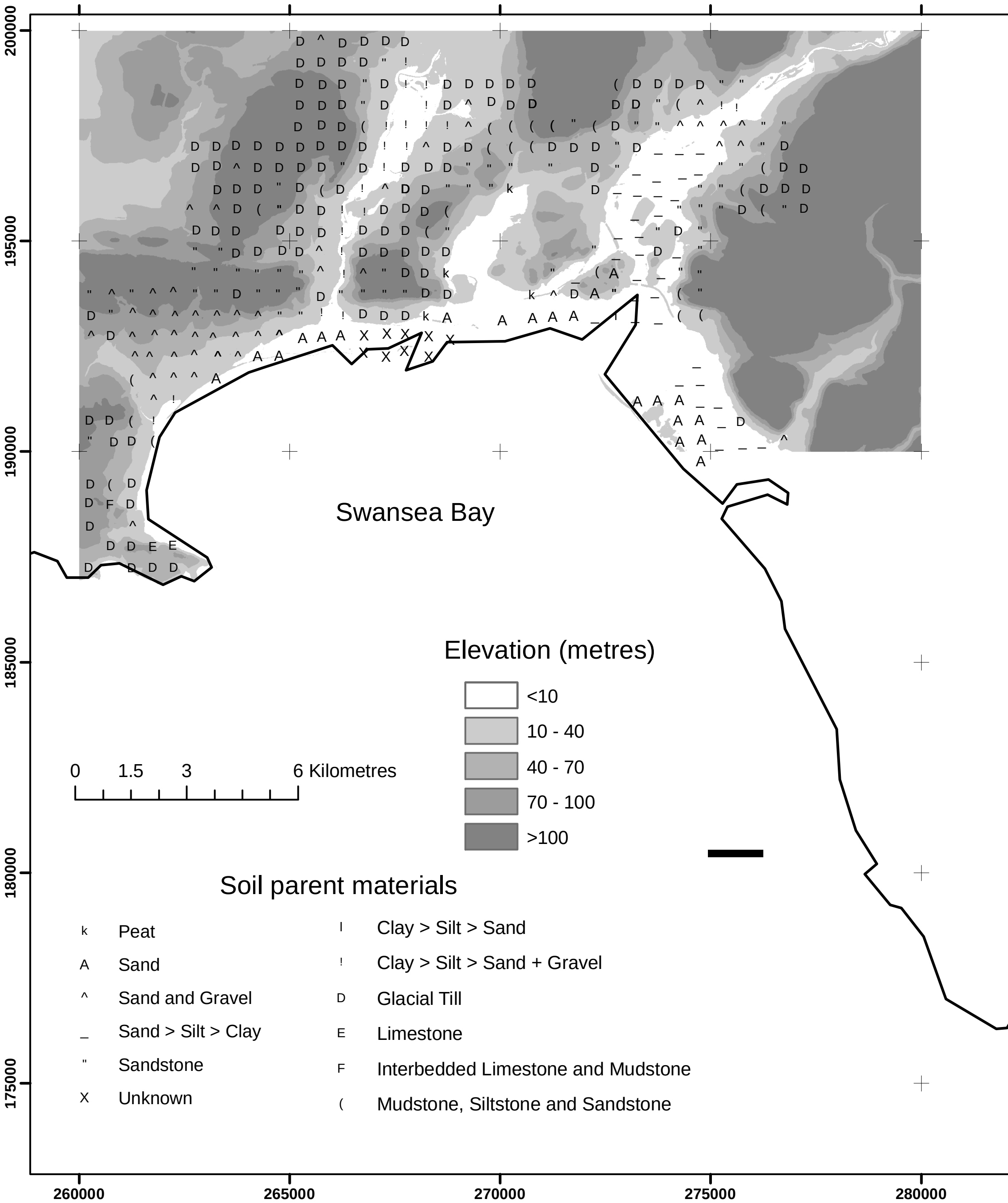

N
0
5 km
Penant Sandstone Formation
Millstone Grit Formation
Old Red Sandstone
Carboniferous Limestone
Glacial sand and gravel
Glacial till
Alluvium
River terrace deposits
Peat
Soil sample points
Swansea

Figure 3:



Swansea Bay
Elevation (metres)
<10
10 - 40
40 - 70
70 - 100
>100
0 1.5 3 6 Kilometres
Soil parent materials
k Peat
A Sand
^ Sand and Gravel
_ Sand > Silt > Clay
" Sandstone
X Unknown
I Clay > Silt > Sand
! Clay > Silt > Sand + Gravel
D Glacial Till
E Limestone
F Interbedded Limestone and Mudstone
(Mudstone, Siltstone and Sandstone
200000
195000
190000
185000
180000
175000
260000
265000
270000
275000
280000

Figure 4:

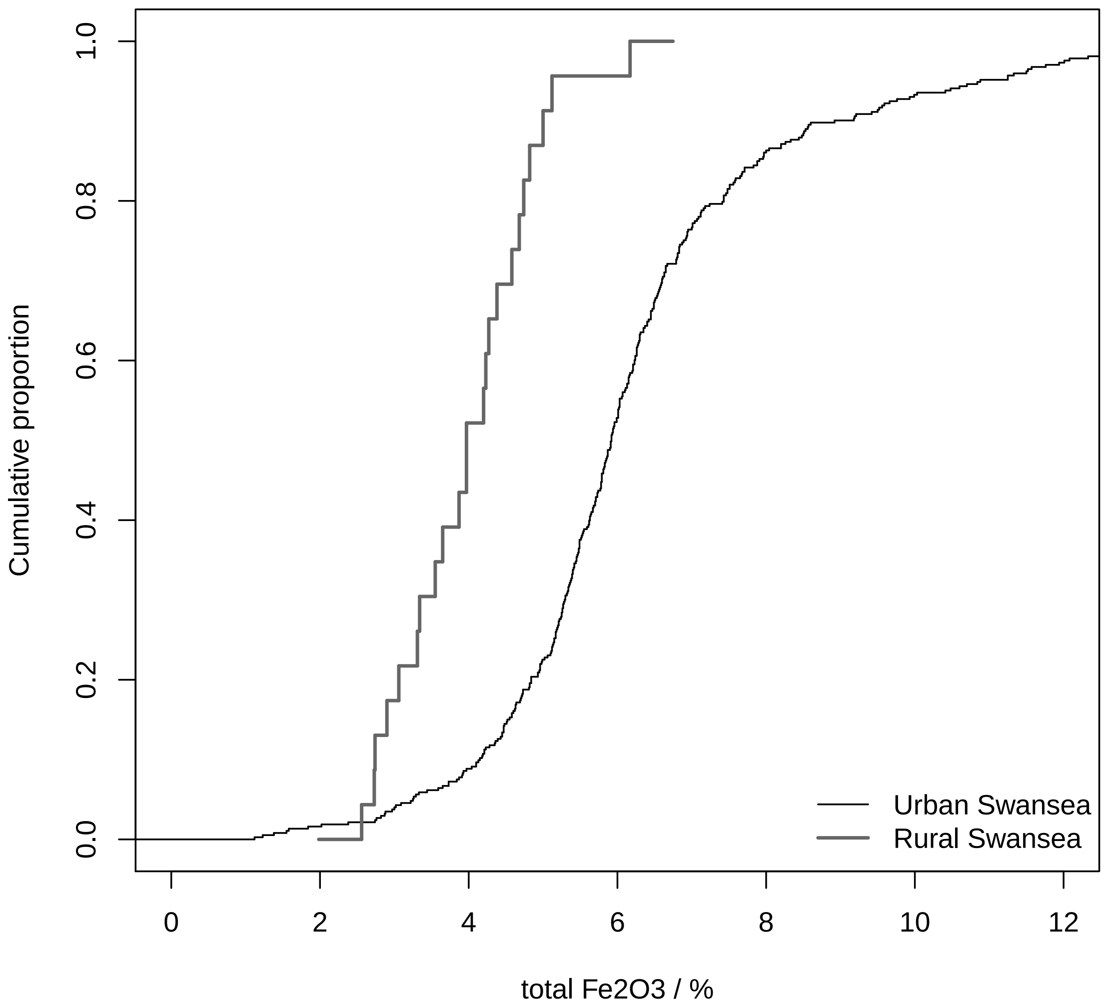
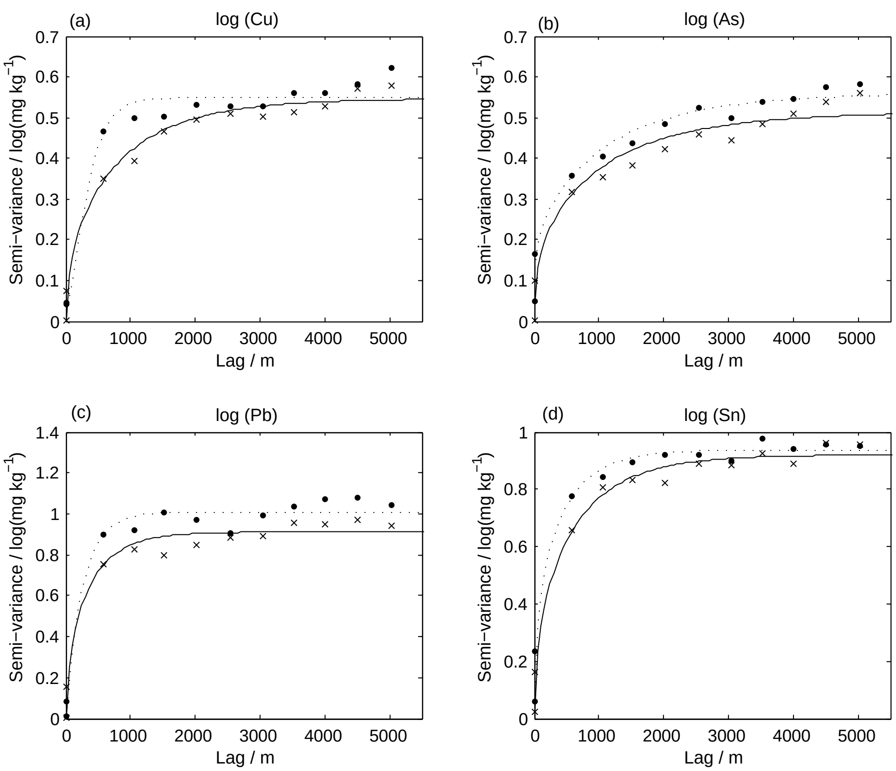
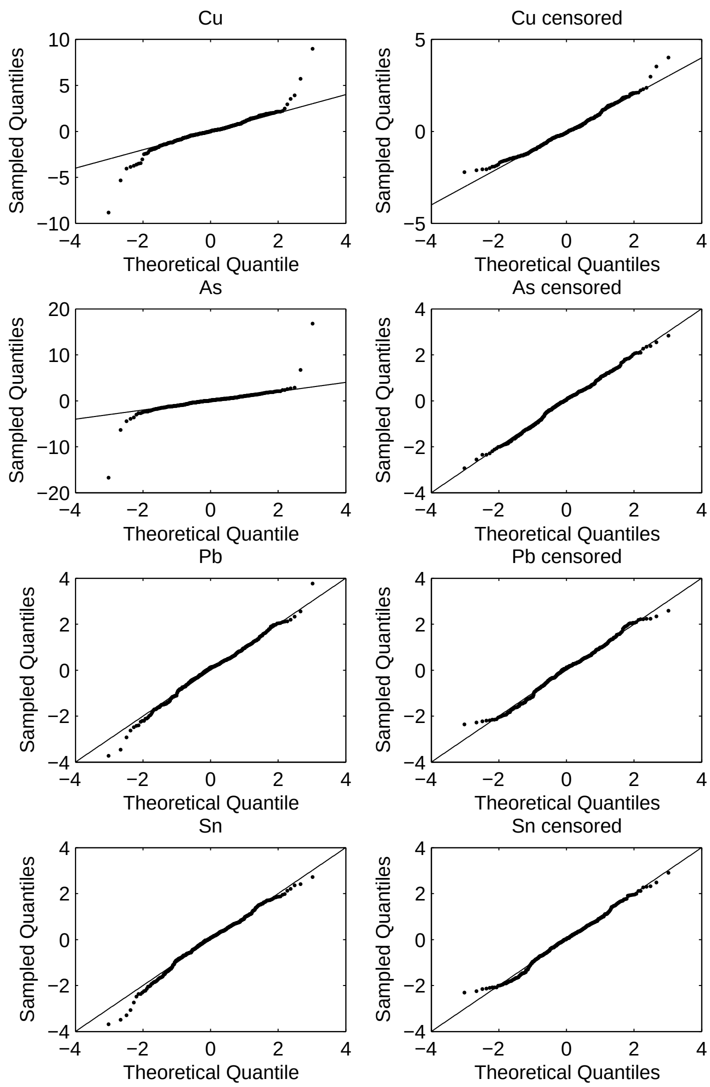

1.0
0.8
0.6
0.4
0.2
0.0
Cumulative proportion
0
2
4
6
8
10
12
total Fe2O3 / %
Urban Swansea
Rural Swansea

Figure 5:



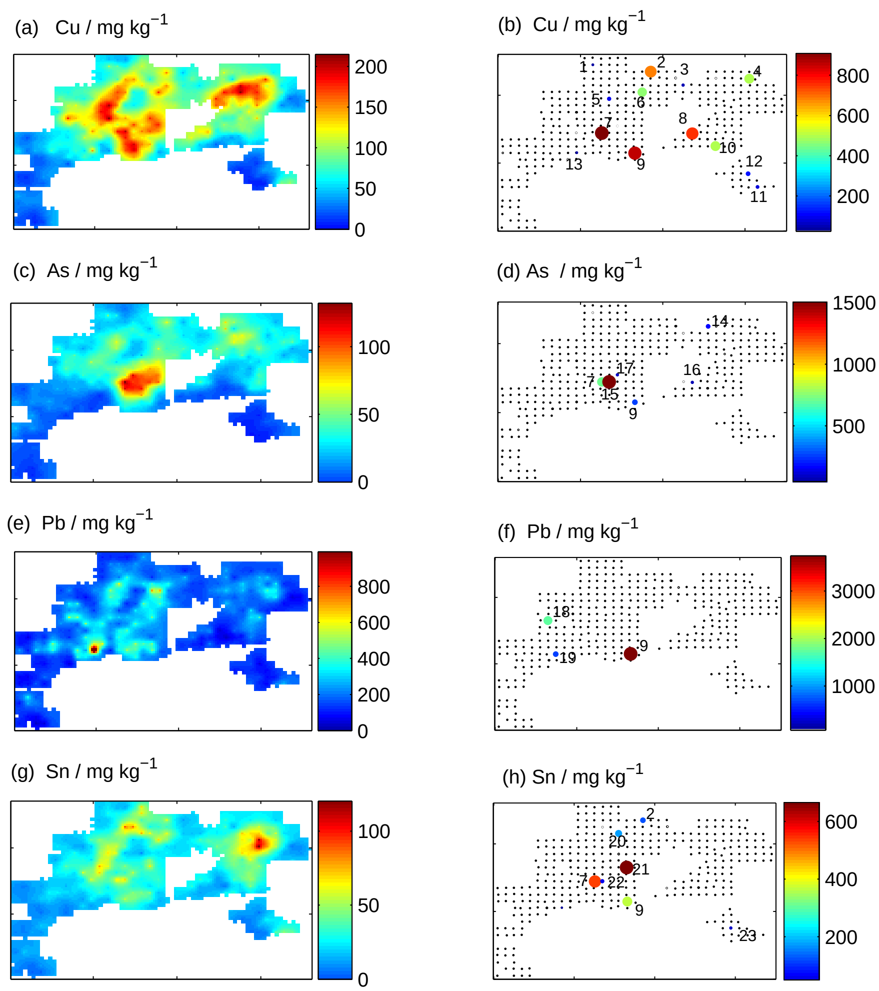
(a)
log (Cu)
Semi−variance / log(mg kg^{-1})
Lag / m
(b)
log (As)
Semi−variance / log(mg kg^{-1})
Lag / m
(c)
log (Pb)
Semi−variance / log(mg kg^{-1})
Lag / m
(d)
log (Sn)
Semi−variance / log(mg kg^{-1})
Lag / m

Figure 6:



Cu
Cu censored
As
As censored
Pb
Pb censored
Sn
Sn censored
Sampled Quantiles
Theoretical Quantile
Theoretical Quantiles

Figure 7:


(a) Cu / mg kg^{-1}
(b) Cu / mg kg^{-1}
(c) As / mg kg^{-1}
(d) As / mg kg^{-1}
(e) Pb / mg kg^{-1}
(f) Pb / mg kg^{-1}
(g) Sn / mg kg^{-1}
(h) Sn / mg kg^{-1}