National-scale estimation of potentially harmful element ambient background concentrations in topsoil using parent material classified soil:stream-sediment relationships

J. D. Appleton[a], B. G. Rawlins[a *] & I. Thornton[b]

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

[b] Imperial College, London, SW7 2AZ, UK

* Corresponding author: B. G. Rawlins, British Geological Survey, Keyworth, Nottingham, NG12 5GG, UK

Tel: +44 (0) 115 936 3140

Fax: +44 (0) 115 936 3200

E-mail address: bgr@bgs.ac.uk (B. G. Rawlins)

Abstract

Regulatory authorities require estimates of ambient background concentrations (ABCs) of potentially harmful elements (PHEs) in topsoil; such data are currently not available in many countries. High resolution soil geochemical data exist for only part of England and Wales (E&W), whilst stream sediment data cover the entire landscape. We present a novel methodology for estimating soil equivalent ABCs for PHEs from high-resolution (HR) stream sediment geochemical data grouped by common parent materials (PM), using arsenic (As) as an example. We use geometric mean (GM) values for local PM groups to investigate different approaches for transforming sediment to soil equivalent concentrations. We use holdout validation to assess: i) the optimum number of samples for calculating local GM values, and ii) the optimum scale at which to group data when using linear regression analysis to estimate GM soil ABCs from local sediment geochemical values. Holdout validation showed that the smallest differences were generally observed when five observations were used to calculate the GM and that these should be grouped over the smallest possible area in order to encompass soils over PMs with elevated GM As concentrations. We estimate and map GM ABCs for arsenic in mineral soil across all of E&W within delineations of PM polygons. Errors for the estimation of soil equivalent GM As ABCs based on sediment data for an independent validation set were of a similar magnitude to those from holdout validation applied to the original data suggesting the approach is robust. Our estimates of soil equivalent ABCs suggest that As exceeds the regulatory threshold used in risk assessments for residential land use (20 mg kg^{-1}) across 16 % of the landscape of E&W. We discuss the applicability of the method for cognate landscapes, and potential refinements.

Keywords: geometric mean, arsenic, parent material, regression, England, Wales

1. Introduction

The ambient background concentration (ABC) of a potentially harmful element (PHE) in topsoil is the sum of the natural (geogenic) and non-natural diffuse components (ISO, 2005, Zhao *et al.*, 2007). In central England, a significant proportion of the landscape has naturally elevated topsoil concentrations of the PHE arsenic (Rawlins *et al.*, 2002) exceeding the Soil Guideline Value (SGV) of 20 mg kg^{-1} for residential land use (DEFRA & EA, 2002a). Regulatory authorities need to know where ABCs are likely to exceed this threshold.

Approaches have been proposed to estimate topsoil ABCs for seven PHEs across parts of the globe using their statistical relationships with total Fe and Mn (Hamon *et al.*, 2004), whilst Zhao *et al.* (2007) did so for several elements across England and Wales (E&W) based on their associations with particle-size fractions. One disadvantage of such approaches is that they require further measurements to be made on samples for which estimates of ABCs are required, and so entail further cost. If high-resolution (HR; sampling intensities greater than 1 sample per 3 km^2) topsoil data were available, it would be possible to provide estimates of ABCs by mapping using some form of local interpolation. Alternatively, where the distribution of elevated ABCs are spatially very complex because they relate to the convoluted outcrops of PHE-enriched soil parent materials (PMs) such as in central England (Palumbo-Roe *et al.*, 2005), we could avoid large errors at these boundaries if we derive estimates of ABCs within delineations of the PM mapping units. This is because soil parent material is the primary control on ABCs in UK topsoil for PHEs including As, Cr and Ni (Rawlins *et al.*, 2003).

At present, this latter mapping approach cannot be used for all of E&W because HR soil geochemical data are only available for around 27 % (area A+B/A+B+C+D in Figure 1) of the landscape; these are soil data from the G-BASE project of the British Geological Survey (Johnson *et al.*, 2005). However, there are high-resolution stream sediment geochemical data for the remainder of the country described by Webb et al. (1978) and Johnson et al. (2005). Preliminary work based on data collected under the G-BASE project showed strong correlations between certain PHEs in soil and stream sediment associated with PM groups across parts of England. Given that the types of PM in this region appear to be representative of much of E&W – comprising a range of geological periods and a significant proportion of Quaternary deposits – we might expect similar relationships to extend nationwide. It may therefore be possible to use the sediment data to estimate topsoil equivalent ABCs for selected PHEs in those areas where soil data are not available. A previous study by Cannon et al. (2004) using a technique of adjusting sediment to soil concentrations reported strong correlations for certain elements across part of Wisconsin, and suggested such an approach could be useful for estimating background values. In contrast, Garrett et al. (2005) were unable to find a routine way of estimating soil concentrations from stream sediment geochemical data in the Upper Coastal Plain of South Carolina, USA due to the complexity of the processes affecting stream sediments during their transformation from soils.

In this paper we present a new methodology for the estimation of topsoil equivalent ABCs for three PHEs (As, Cr, Ni) using HR stream sediment data, and demonstrate its application to soil As. We establish statistical relationships between geometric mean values of soil and stream sediment PHE concentrations grouped by PM, and use these to estimate mineral topsoil equivalent concentrations of As, which we map within delineations of the PM polygons. We use an independent dataset to demonstrate the robustness of our approach. We present the first national scale map of topsoil As ABCs (based on geometric mean values for delineations of PM polygons) resulting from the application of our methodology. We also show how these data can be used to estimate the proportion of samples exceeding a threshold used in regulation related to contaminated land assessments. Finally, we discuss the uncertainties associated with our methodology, its wider implications and potential refinements.

2. Exploratory data analyses for estimating soil equivalent ABCs

We require a method to transform the available HR stream sediment and deeper soil geochemical data for E&W (Figure 1) into topsoil equivalent ABCs. The HR geochemical survey data used in this study, including analytical methods, sampling density and dates are summarised in Table 1 and Figure 1. Wolfson stream sediment samples were taken from small tributaries with catchments that rarely exceeded 5-10 km^2 whilst the GBASE stream sediment samples were collected from small, first or second order, streams to give an average sampling density of one sample per 1.5 to 2 km^2. Total element concentrations were determined so these are compatible with the SGV regulatory thresholds for England and Wales. The data can be separated into four, spatially overlapping combinations of topsoil, subsoil and stream sediments from the GBASE survey, and Wolfson stream sediment survey (regions A to D; Figure 1). Topsoil (0-15cm depth) geochemical data from the GBASE survey were available in region A. Deeper soil (35-50cm depth) geochemical data from GBASE were available in regions A+B. Stream sediment geochemical data from the GBASE survey were available in regions A+B+C covering large areas of north and central England. Finally, stream sediment geochemical data are available for all of E&W (A+B+C+D) from the Wolfson Atlas.

We undertook two sets of exploratory analyses. First, we created scatterplots of GM concentrations in topsoil versus deeper soil for As, Cr and Ni grouped by soil PM for area A (Figure 2) and fitted linear regressions to them using least squares (see Table 2). These highlight very strong linear relationships – the slopes are all close to one. The slope of the linear regression between GM topsoil and deeper soil As (shown in Figure 2a) is 1.01, so we felt justified in treating PM grouped topsoil and subsoil As values as equivalent. For Ni and Cr, we would need to apply a linear transformation to estimate GM concentrations based on samples grouped by PM. Higher Cr and Ni concentrations in the <150 µm fraction of deeper soils compared with the <2mm fraction of topsoils is to be expected whilst the approximately equivalent As concentrations in the two soil sampling media requires further investigation. Second, we assessed the significance of PM in determining the spatial distribution of As, Cr and Ni in the large GBASE dataset for deeper soils and GBASE stream sediments for area A+B (Figure 1). The results, summarised in Table 3, demonstrate the primary importance of PM in determining the concentrations of these elements in both soil and stream sediment, with the variance accounted for ranging from 20 to 43%. There were strong correlations between geometric mean (GM) PHE concentrations in soil and stream sediment when the data were grouped by PM. We therefore considered that it was justified to investigate whether we could estimate soil equivalent ABCs using stream sediment PHE concentrations in areas C+D (Figure 1) where no soil geochemical data were available.

Given that the samples are grouped by PM, we required a statistical measure of location to express the ABC. We examined features of the statistical distributions of As, Cr and Ni for areas A+B (Figure 1) where we can compare soil and stream sediment geochemical data (see Table 3). All of the variates had large positive skewness coefficients for each of the PHEs. After transforming the data by taking natural logarithms the skewness coefficients were generally in the range [-1,1] suggesting that the majority of the original data were approximately log-normally distributed. Traditional measures of statistical location (mean) and scale (standard deviation) are biased when applied to skewed distributions. To overcome this we used the geometric mean (GM) and geometric standard deviation (GSD) to establish statistical relationships between variates using the original, untransformed data. Our estimates of ABCs are GM values for PM groups, which are similar to the medians in each distribution. The latter was the parameter proposed for estimating ABCs by ISO (2005). However, GM provides a better estimate of ABC than median when calculating ABCs from small numbers of samples, especially for PMs with relatively high arsenic concentrations.

In the next section we describe how we evaluated different features of an approach for the conversion of stream sediment to soil equivalent ABCs using statistical relationships based on data grouped by PM. Specifically, we use holdout validation to test: i) different scales for grouping soil and stream sediment geochemical data by PM and, ii) the optimum number of neighbouring samples required to calculate GM concentrations . We then demonstrate how linear regression relationships between sediment and soil for common PM groups can be used to estimate ABCs in topsoil in areas C+D in Figure 1. Prior to this we transformed the Wolfson stream sediment in southern England (area D; Figure 1) to their G-BASE equivalents using linear quantile transformation.

3. Statistical and mapping methods

Below we describe the detailed methodology for transformation of the available stream sediment As data into soil equivalent ABCs with reference to a sequence of steps shown in a flowchart (Figure 3).

3.1. Linear quantile transformation (steps 1 and 2)

To transform the Wolfson data to the G-BASE sediment data we used linear quantile transformation (Daneshfar and Cameron, 1998, Darnley et al., 1995, Heyde, 1986). Here we briefly summarise the theory of quantile transformation. If F(x) is some distribution function on the real line, and G(x) is another, and we have a random variable Y with distribution function G, we want to create a random function X with distribution function F, so that the difference $|X - Y|$ is as small as possible. This can be achieved if we define the variable ξ which is uniformly distributed on [0, 1] by $\xi = G(Y\ )$, and then set $X = F-1(\xi)$. In our case the random variable X are concentrations of a PHE in Wolfson sediment samples, and variable Y are concentrations for the same PHE in the G-BASE samples for the same geographic area

(A+B+C in Figure 1). We can then fit a linear regression using least squares for a series of quantiles (e.g. p = 0.1, 0.2, . . . , 0.9) between the target distribution (Yp) and the source (Xp). We can apply the regression to estimate the concentrations of Y (G-BASE) from X (Wolfson). We assume that the sampling method is unbiased in both cases, and that that X and Y are related by a positive, linear scaling.

3.2. Parent material polygon delineations as geochemical mapping units (Step 3)

In this study, we defined PM classes based on the concatenation of separate codes for the underlying bedrock and any superficial deposits present (see Figure 4). The codes are generally derived from digital versions of the 1:50,000 maps of bedrock geology and superficial deposits for E&W, part of DigMap GB (British Geological Survey, 2006). Initially a total of *ca*. 1.9 million individually delineated polygons were created in ArcMap^{TM} GIS (ESRI) by separating unioned bedrock and superficial geology polygons using a 1-km grid aligned to the British National Grid (see Figure 4). There was frequently more than one polygon of a PM within a 1-km grid square. In such cases, the average centroid for a PM in a 1-km grid was calculated from the centroids of the individual polygons of that PM within the 1-km grid square (see bottom right 1km grid square in Figure 4). There are approximately 650,000 average 1km-PM centroids across E&W and these are used to estimate the GM concentrations for the delineations of the 1km-PM polygons used as geochemical mapping units in this study. The use of the average centroids, rather than individual 1km-PM polygon centroids reduced geochemical mapping computation time by approximately 65%. We used a spatial join procedure (ArcMap^{TM} software (ESRI)) to link the geochemical sampling sites to their PM code. The GM ABC for each PM in a 1-km grid square is calculated from the *n* geochemical samples located on the same PM that are nearest to the average 1km-PM centroid. The optimum number (*n*) of samples for estimation of GM was determined using holdout validation, as explained below.

3.3. Holdout validation to determine optimum number of samples for estimating sediment and soil GM (step 4)

We wished to assess the optimum number of local sediment or soil samples with the same PM code for calculating GM values for each 1km-PM polygon. This is likely to vary due to a range of factors including the spatial distribution of sampling locations, the size and shape of PM polygons, and drainage pattern. We used a script written in the GIS package ArcView^{TM} (ESRI) which identified for each average 1km-PM polygon centroid, the nearest ‘n’ (1, 2, 3, 4 , 5, 7, 9, 11, 15, 20 and 30) soil or sediment sampling sites located on the same PM. This script returned from the *n* nearest sediment or soil samples for each PM class: i) the GM As concentration, ii) the inverse distance weighted value of their natural log transformed As concentrations (on the same scale as the GM), and iii) the distance to the furthest of the *n* sediment or soil samples. We used a holdout validation procedure in which a random subset of 10% of the sediment or soil sites were removed, using the remaining 90% to estimate GM (GM_{est}) values at the sites of the former from the *n* (1, 2, 3, 4, 5, 7, 9, 11, 15, 20 and 30) nearest neighbouring sediment sites on the same PM. In the case of GM values, we calculated the Mean Squared Deviation (MSD) between the estimated (GM) As at the site and the measured As at each of the sites in the random 10% subset:

$$\mathrm{MSD} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}\left(\mathrm{GM}_{\mathrm{est}} - \mathrm{measured}\right)^{2} \qquad \text{(Equation 1)}$$

We also calculated MSD’s between the natural log transformed measured soil As and its estimate based on the inverse distance weighted value based on the same log transformed stream sediment data.

We undertook this analysis using soil data for two similar Jurassic ironstone units (Northampton Sand Formation (INONS) and the Marlstone Rock (MRB)) and also for the Upper Lias (ULI) which is the only other geological unit in the area with substantial lateral variation in arsenic concentration. For the stream sediment data we undertook the same holdout validation for three randomly selected 10% subsets of one PM group characterised by substantial lateral variations in arsenic concentrations (the Lower Silurian mudstone dominant sedimentary rocks of Central Wales (SLLA-MDMIX)). We used these results to determine an appropriate number of samples to calculate GM PHE values for both soil (areas A+B; Figure 1) and stream sediments (areas C+D).

3.4. Grouping of samples by PM class at different scales (step 5)

We wished to use regression to estimate GM soil As concentrations (predictand) for unique PM polygons using their local GM sediment concentrations (predictor). We needed to define the minimum number of soil and sediment samples required for the calculation of a GM value because if this is based on too few samples, the GM will be imprecise. We chose to limit the regression analysis to include only those PM groups with more than four samples. Exploratory analysis showed that when the data were grouped by 10 km squares across central England there were relatively few (<4) samples over several of the thin, iron-rich PM outcrops which have elevated PHE concentrations. In the 10-km grid square illustrated in Figure 5, for example, there are only two sediment samples located on the Marlstone Rock Formation. Excluding these PM groups from the regression analysis could introduce bias. By grouping at different spatial scales (e.g. 1 km^2, 25 km^2, 100 km^2 and greater) we can investigate the influence of grouping scale on the regression models. We grouped sediment and soil samples by their PM class using three approaches. First, using all the average 1-km PM polygon centroids, we identified the nearest five sediment samples located on the same PM and calculated GM As. No value was reported where less than 5 samples are available for a PM.

This approach returned 29 416 comparisons of the local sediment (GM_{sed}) and soil (GM_{soil}) GM As concentrations with common PM classes. Second, by averaging over 25 km^2 grid squares the GM values for PM codes derived from the first approach and comparing these to common PM soil GM As values (n=4025). Third, using a nested-scale approach in which groups of *n*>4 samples with the same PM code were identified in order of increasing scale within: i) 5 km grid squares (25 km^2), ii) 10 km grid squares (100 km^2), iii) geological map sheet (approx 550 km^2), and iv) 100 km grid square (1000 km^2). If insufficient samples were present at the smaller scale, the next greater scale was used. This resulted in 1188 paired PM soil and stream sediment GM mean As concentrations. By adopting this latter approach, we ensure that iron-rich PM groups with elevated PHE concentrations are included in the regression analysis. Regression equations were validated by calculating the Mean Squared Deviation (MSD) between the measured As at each soil sample site and the estimated GM soil As (GM $soil_{est}$) calculated from the nearest five sediment values on the same PM (Equation 2):

$$\mathrm{MSD} = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \left(\mathrm{GM\ soil}_{\mathrm{est}} - \mathrm{measured}\right)^{2} \qquad \text{(Equation 2)}$$

3.5. Regression and estimation of confidence intervals (steps 6 and 7)

We investigated both linear and polynomial regression relationships; the latter has the form:

$$y = \alpha + \beta_1 x + \beta_2 x^2 + \varepsilon \qquad \text{(Equation 3)}$$

in which the sediment GM As concentration for each PM is the explanatory variable (x), with which we wish to predict the equivalent GM PHE for soil (y) for the same PM code, with ε representing any unexplained variation. Second order polynomial regression models were used after it was found that these were more suited to the transformation of sediment to equivalent soil concentrations than first order polynomial regression, particularly at high concentrations above the SGV. Least trimmed squares approaches had the same limitations as the first order polynomial models and so we used the method of least squares.

We assessed the impact of grouping sample locations based on different scales and the numbers of samples in each group for their impact on: i) the regression relationships between soil and stream-sediment PHE concentrations, and ii) holdout validation statistics for estimation based on these regression relationships. The holdout validation statistics used were the Mean Squared Deviation (Equation 4) between PM grouped GM soil$_{\mathrm{est}}$ and GMsoil, the root-mean-squared deviation (RMSD; Equation 5), and bias (Equation 6):

$$MSD = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \left(\mathrm{GMsoil} - \mathrm{GMsoil}_{\mathrm{est}}\right)^{2} \qquad \text{(Equation 4)}$$

$$RMSD = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \left(\mathrm{GMsoil} - \mathrm{GMsoil}_{est}\right)^{2}} \qquad \text{(Equation 5)}$$

$$Bias = \sum_{n} \left(\mathrm{GMsoil} - \mathrm{GMsoil}_{est}\right)^{2} / N \qquad \text{(Equation 6).}$$

In establishing a regression relationship between GM As concentrations for PM groups (explanatory variable, (x_i)) and the proportion of samples exceeding the regulatory threshold (y_i ; the predictand) of 20 mg kg $^{-1}$, we also considered non-linear relationships of the form:

$$y_i = \alpha + \beta \, \rho^{x_i} + \varepsilon_i \qquad \text{(Equation 7)}$$

where α , β and ρ are estimated parameters from a curve fitting procedure.

We calculated the 95% confidence limits around the estimated GM As concentration for each polygon centroid using the mean ($\bar{x}$) and standard deviation (*s*) of their log-transformed values. The confidence limits were calculated on the log transformed scale:

$$\bar{x} - t_f \frac{s}{\sqrt{N}} \quad \text{and} \quad \bar{x} + t_f \frac{s}{\sqrt{N}} \qquad \text{(Equation 8)}$$

where t_f refers to the value for the Students *t* distribution at the 2.5% significance level for *N*-1 degrees of freedom (2.776) and *s* is the standard deviation of the nearest 5 soil or sediment (log-transformed) values. We estimated the confidence interval for soil and sediment polygon centroids separately. The confidence limits were then back-transformed to the measurement units of the original scale.

Finally, we used the GM As concentration ($\bar{x}$ Y) and the geometric standard deviation ($s^2{}_Y$) for PM grouped data to estimate the proportion of samples exceeding the regulatory threshold (*z*) using the formula for the standard normal distribution:

$$f(z; \bar{x}_Y, s^2{}_Y) = \frac{1}{\sqrt{2\pi s^2{}_Y}} e^{-\left(\frac{(z - \bar{x}_Y)^2}{2 s^2{}_Y{}^2}\right)} \qquad \text{(Equation 9).}$$

The regression model between GM As and the estimated proportion of samples exceeding the SGV (%>SGV), derived from PM grouped soil data, was used to estimate the %>SGV for individual 1km-PM polygons for which GMs were calculated from the nearest 5 soil samples located on the same PM (or the soil equivalent GM As estimated from sediment data as described above).

3.6. Independent validation for estimation of topsoil As ABCs using sediment data (step 8)

To independently validate the conversion of sediment to soil equivalent As ABCs we used analyses by X-ray Fluorescence Spectrometry from a set of 11,335 topsoil samples collected by the British Geological Survey between 1972 and 1997. The samples were collected in areas with mineralisation potential, predominantly in Devon and Cornwall, Pembrokeshire, the Lake District and Northumberland (Figure 1). This is reflected by the elevated values of parameters from the statistical distribution of the variate (mg kg^{-1}), mean (78), median (25), GM (32), maximum (8647) and strong positive skewness (11.7). These samples are located within 414 1km-PM polygons, with a minimum of 4 samples in a 1km-PM polygon. For data grouped into ranges of GM As we then calculated the MSD between (i) GM As for the samples within each 1km-PM polygon and (ii) GM soil As for the 1km-PM polygon estimated from stream sediment data using the linear regression based on 1188 paired PM soil and stream sediment GM mean As concentrations, as explained above. We also calculated the bias for the data as a whole.

4. Results and their interpretation

4.1. Linear quantile transformation

We calculated percentiles (*p*=10, 20,…, 90, 95 and 99) for both the Wolfson and G-BASE stream sediment As data for area A+B+C (Table 4) and fitted a linear regression through them using least squares (Figure 6). The regression accounted for 99.2 % of the variance, with an intercept (α) of -3.1 and slope (β) of 1.42 based on the eleven paired percentiles. Weighted linear regression models, as used by Daneshfar and Cameron (1998), produce almost identical results over the same percentile range. For comparison we also plotted the GM As concentrations for PM groups with greater than eight samples for each of the datasets. A linear regression of these points (not shown) was very similar to that for the percentiles.

All the paired quantiles and the majority of the PM GMs plot above the 1:1 line indicating that in general As concentrations are higher in the finer (<150µm) G-BASE sediments than the slightly coarser (<177µm) Wolfson samples. This conforms with geochemical theory in which some trace elements are enriched in finer-grained samples (Plant, 1971), due to their associations with iron and manganese oxy-hydroxides. The regression equation described above for areas A+B+C was then applied to the Wolfson sediment data in area D to create a continuous map of As in stream sediment for areas C+D. These data were then converted to soil equivalent concentrations using the methods that follow.

4.2. Holdout validation: optimum number of samples for estimating GM

The MSD is relatively constant when the number of neighbouring soil samples used to calculate GM exceeds 4 for those PMs which exhibit relatively little lateral variation in As whereas the variation in MSD with 'n' is strongest for those PMs which exhibit marked lateral variation in GM As (e.g. Northampton Sand Formation (INONS) and the Marlstone Rock Formation (MRB) ironstone units). In order to increase the reliability of the MSD tests, soil data for the two similar Jurassic ironstone units (INONS and MRB) were grouped together to produce a subgroup of 320 soil samples. The only other PM in the area with substantial lateral variation in arsenic is the Upper Lias (ULI, n = 315 soil samples). Average results of three holdout validation calculations for these two groups of soils are presented in Figure 7a. In general, the differences (MSD) between estimated and actual GM As values decrease as the number of neighbouring samples used to calculated the GM increases from one to five. The optimum number of samples (smallest MSD) for estimating GM soil arsenic for 1km-PM polygons is between 5 and 9.

The results for three sets of holdout validation for sediments from the SLLA-MDMIX (Silurian Llandovery mudstone-dominant sedimentary rocks of Central Wales with no superficial cover) PM are presented in Figure 7b. In subset 1, there is a gradual decrease of MSD as the *n* samples used to determine the GM increases from 1 to 7 and then an increase in MSD from 7 to 25. For subsets 2 and 3, there is little change in MSD when the number of samples is 5 or more whilst the MSD is higher when *n* = 1 or 2. Selecting the optimum number of samples is a balance between the number which gives the smallest MSD whilst ensuring that significant local variations in arsenic concentrations are diminished by using too many samples. The optimum number of samples for estimating arsenic for the centroids of 1km-parent material polygons is between 4 and 7 for sediments, so we decided to calculate GMs from the 5 nearest samples for both sediments and soils.

4.3. Linear regression, holdout validation and mapping soil As

The second order polynomial model (Equation 3) fitted to the three groups of GMsoil and GMsoil$_{est}$ data are shown in Figure 8a-c and the regression coefficients in Table 5. The holdout validation statistics are shown in Table 6. The regression based on grouping at a range of scales (Figure 8c) returns higher estimates of GM soil As in the upper range (>3 log normal transformed As ~ 20 mg kg^{-1}) because there are more As enriched PM types with *n*>4 samples included, reflected in the regression equation plotting closer to the 1:1 line than those in Figures 8a and 8b. The standard error of the estimate for the nested-scale for grouping PM groups was also smaller than for grouping using the two other scales (Table 5).

The smallest MSD (Equation 2) values at the lower As concentrations (0 - 20 mg kg^{-1}) were those for regressions based on the 1-km polygons and 5-km averaging of 1-km polygons (Figures 8a-b), with slightly higher values for the regression derived from the nested-scale approach. However, the MSD values are significantly smaller above 20 mg kg^{-1} As for the regression in which the nested approach was used based on PM groups across four different scales (Table 6). Given the importance of accurately estimating GM As values in soil with elevated ABCs of trace elements, we selected the nested-scale approach and its regression equation (Fig 8c) for the conversion of PM grouped sediment to soil equivalent As concentrations. The MSD and RMSD (Equation 5) values for classes with increasing distances for GM versus an IDW interpolation of As concentrations of the nearest five sediment samples (Table 7) show that: i) the former has smaller errors, ii) that the variation in these errors is smaller, and iii) estimation errors increase with maximum distance to the furthest sample. Therefore we chose to base the estimation of soil equivalent As on the GM of the nearest five sediment samples for common 1-km PM polygons rather than the IDW estimate.

The final map of mineral topsoil As ABCs was produced by combining the GM As concentrations for polygons in areas A+B (Figure 1) using the deeper soil G-BASE data and the topsoil equivalent estimates for GM As resulting from application of steps 1-6 (Figure 3) to the stream sediment data for areas C+D (Figure 1). The map of topsoil estimated ABCs (Figure 9) shows that As background concentrations are likely to exceed the regulatory threshold over 16% of E&W. When we include those areas that are equal to the regulatory threshold the proportion increases significantly to 25%. The fine resolution of the soil equivalent concentrations shown is due to the in excess of 1.9 million individual 1km-PM polygons across E&W. The boundaries between the categories of soil concentrations are sharp, reflecting the delineations of the PM polygons. This is a noticeable difference to those maps based on interpolation of data at discrete sampling locations. The largest areas with the highest concentrations (>30 mg kg^{-1}) occur in the English Lake District, western Wales and south-west England. No estimates can be made for ABCs in the greater London area because no stream sediment or soil data are available to date.

The estimated 95% confidence intervals for each of the concentration classes shown on Figure 9 are presented in Table 8; confidence intervals are presented for ABCs estimated from both soil and sediment values. The confidence intervals become wider as the GM As ABC increases. Also, the confidence intervals for PM polygons based on sediment values are greater than those for each of the soil concentration ranges. Hence, the greatest uncertainties for ABCs shown in Figure 9 are for those areas of Wales, north-west England and south-west England where the large estimated ABC As values (>30 mg kg^{-1}) are based on stream sediment data.

4.4. Independent validation

The MSD and bias values calculated from the independent validation dataset are shown in Table 6. The validation data were sited over mineralised areas of E&W where without prior information it would be difficult to estimate soil As ABCs. This is reflected in the MSD values, which are somewhat larger (0.25-2.08) than those for the holdout validation based on the nested-scale regression (0.1-1.79). The bias of the estimates is also somewhat larger; 0.19 for the independent validation compared to 0.03 after application of the nested-scale regression. However, these independent validation data demonstrate the methodology is sufficiently robust to be used for the estimation of ABCs across the landscape of E&W for those elements which have similar geological and geochemical controls to As. The methodology has not been assessed for elements that dominantly occur in resistate minerals, such as Zr, Sn, or W, especially in areas of strong relief.

4.5. Probability of exceeding the As regulatory threshold

The least squares fit of the non-linear regression relationship between GM As and the proportion of samples in each PM group exceeding 20 mg kg^{-1}, fitted using the CURVEFIT directive in Genstat (Payne, 2002) is plotted in Figure 10. The coefficients from Equation 7 were: α (99.7), β (-171.9) and ρ (0.944) with the regression capturing 92.3% of the variance, with a standard error of 6.26. So using the GM As concentration for any soil equivalent PM polygon and the normal distribution (Equation 7) we can apply this regression equation to estimate the proportion of samples which are likely to exceed the regulatory threshold of 20 mg kg^{-1}.

5. Discussion

We have presented and applied a methodology which employs stream sediment values to estimate ABCs of As in mineral topsoil across E&W, based on common PM groups. Our analysis suggests that 16% of the landscape of E&W has As ABCs exceeding the threshold (i.e. 20 mg kg^{-1}) of the first tier of the risk assessment adopted by regulatory authorities for residential land use. Although topsoil As ABCs above this threshold does not in itself imply a significant health risk, it does show that more frequent and complex contaminated-ground risk assessments will likely be needed for much of the landscape of E&W. These spatially referenced data can be used to assess the probability that the SGV will be exceeded at a particular site, and whether elevated concentrations of arsenic observed in site investigations may be attributable to geogenic sources or whether it is possible that the observed concentrations may have been influenced by anthropogenic factors. From our preliminary analysis (Figure 3), we believe this approach could be extended to include Cr and Ni. The latter may be of particular significance because it exhibits geogenically elevated concentrations (McGrath and Loveland, 1992), exceeding the SGV of 50 mg kg^{-1} (DEFRA & EA, 2002b) across parts of E&W.

Whether this methodology could be applied more widely depends on a number of factors, but principally on the correlations between PHEs in soil and stream sediments across large areas. It will be most applicable where soils are relatively young, such as the recently glaciated areas of northern Europe, and cognate landscapes. Clearly there is a need for existing, HR stream sediment geochemical data; this is often available from mineral exploration studies. For example, the National Uranium Resource Evaluation Programme in the USA has around 400,000 stream sediment samples (Bolivar, 1980), compared with geochemistry for only 1323 soil samples available nationally (Boerngen and Shacklette, 1981).

We recognise there may be theoretical objections to our approach. First, stream sediments represent the weathered material transported from an entire catchment which may comprise a number of geological sources with differing geochemical compositions. When eroded and transported along the stream network, the geochemical composition of mixed provenance sediment may be quite different from the chemistry of the underlying geological substrate at any particular point in the stream, and also the soils derived from its PM. Second, in-stream geochemical processes tend to increase the concentrations of PHEs due to their strong adsorption to, or co-precipitation with, iron and manganese oxides which commonly coat stream sediments (Nichol *et al.*, 1967), leading to potentially significant bias. This may be greatest in upland areas of E&W, where secondary precipitation may be enhanced in acidic streams draining organic rich, peat soils. The potential for overestimation of ABCs in soil based on stream sediment in these environments requires further investigation especially with respect to the concentration of chemical elements and mineral species in the fluvial environment (Garrett et al., 2005).

It may also be possible to enhance our methodology. For example, the use of weighted linear regression may improve the transformation of sediment to soil equivalent concentrations, and this could be tested using holdout validation. Second, we could explore the scale-dependent correlation of sediment and soil concentrations using a geostatistical approach. This might indicate that different scales of generalization may be more appropriate than the parent material unit. As the variables are not collocated, it would be necessary to use the pseudo cross-variogram (Myers, 1991). To these we could fit coregionalization models to compute correlations between the mean values of the variables within blocks of different size (the inter-block correlation; (Pringle and Lark, 2007)). This could be used to assess whether there are advantages in using grouping at scales different to those we have used here.

6. Acknowledgements

This paper is published with the permission of the Director of the British Geological Survey (Natural Environment Research Council). We would like to thank Murray Lark from Rothamsted Research (UK) who provided advice on some aspects of the statistical analysis used in this study. The authors would like to thank all BGS staff involved in the collection and analysis of samples from the GBASE survey. R.G. Garrett and W.F. Cannon provided very useful reviews and are thanked for their comments and suggestions.

7. References

Boerngen, J.G., Shacklette, H.T., 1981. Chemical analyses of soils and other surficial materials of the conterminous United States. United States Geological Survey Open-File Report 81-197. USGS, Denver.

Bolivar, S.L., 1980, Los Alamos, N.M., 1980 An overview of the National Uranium Resource Evaluation Hydrogeochemical and Stream Sediment Reconnaissance Program. U.S. Department of Energy, Grand Junction, Colorado, GJBX-220(80), pp. 24.

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.

Cannon, W.F., Woodruff, L.G., Pimley, S., 2004. Some statistical relationships between stream sediment and soil geochemistry in northwestern Wisconsin - can stream sediment compositions be used to predict compositions of soils in glaciated terranes? J. Geochem. Explor. 81, 29-46.

Daneshfar, B., Cameron, E., 1998. Levelling geochemical data between map sheets. J. Geochem. Explor. 63, 189-201.

Darnley, A.G., Björklund, A., Bølviken, B., Gustavsson, N., Koval, P.V., Plant, J.A., Steenfelt, A., Tuachid, M., Xuejing, X., Garrett, R.G. and Hall, G.E.M. 1995. A Global Geochemical Database for Environmental and Resource Management: Recommendations for International Geochemical Mapping, Final Report of IGCP Project 259. (Paris: UNESCO)

DEFRA & EA, 2002a. Soil Guideline Values for Arsenic Contamination. Department of the Environment Food and Rural Affairs and the Environment Agency, Bristol, pp. 14.

DEFRA & EA, 2002b. Soil Guideline Values for Nickel Contamination. Department of the Environment Food and Rural Affairs and the Environment Agency, Bristol, pp. 20.

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. Glob. Biogeochem. Cycle 18, 1-6.

Garrett, R .G., Drew, L.J. and Sutphin, D.M. 2005. Estimated soil geochemistry from stream sediment geochemistry. In: GIS and Spatial Analysis: Proceeding of 2005 Annual Conference of the International Association for Mathematical Geology (IAMG), 1, 452-457.

Heyde, C.C., 1986. Quantile transformation methods. In: Kotz, S., Johnson, N.L., Read, C.B. (Eds.), Encyclopedia of Statistical Sciences Vol. 7. John Wiley & Sons, New York.

ISO, 2005. Soil Quality: Guidance on the determination of background values. International Organisation for Standardisation. ISO 19258:2005

Johnson, C.C., Breward, N., Ander, E.L., Ault, L., 2005. G-BASE: Baseline geochemical mapping of Great Britain and Northern Ireland. Geochemistry: Exploration-Environment-Analysis 5, 1-13.

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

Myers, D.E., 1991. Pseudo-cross variograms, positive-definiteness and cokriging. Math. Geol. 23, 805-816.

Nichol, I., Horsnail, R. F., Webb, J. S. 1967. Geochemical patterns in stream sediment related to precipitation of manganese oxides. Trans. Inst. Min Metall., Lon., Series B 76, B113-115.

Palumbo-Roe, B., Cave, M.R., Klinck, B.A., Wragg, J., Taylor, H., O'Donnell, K., Shaw, R.A., 2005. Bioaccessibility of arsenic in soils developed over Jurassic ironstones in eastern England. Environ. Geochem. Hlth. 27, 121-130.

Payne, R.W., 2002. The Guide to Genstat: Part 2 Statistics. VSN International, Oxford.

Plant, J.A., 1971. Orientation studies on stream sediment sampling for a regional geochemical survey in northern Scotland. Trans. Inst. Min, Metall., Lon., Series B, B234-345.

Pringle, M.J., Lark, R.M., 2007. Scale- and location-dependent correlations of soil strength and the yield of wheat. Soil Till. Res. 95, 47-60.

Rawlins, B.G., Lister, T.R., Cave, M., 2002. Arsenic in UK soils: reassessing the risks. Proc. Inst. Civil Eng. 150, 187-190.

Rawlins, B.G., Webster, R., Lister, T.R., 2003. The influence of parent material on top soil geochemistry in eastern England. Earth Surf. Proc. Land. 28, 1389-1409.

Webb, J.S., Thornton, I., Howarth, R.J., Thomson, M., Loewnstein, P., 1978. The Wolfsen Geochemical Atlas of England and Wales. Clarendon Press, Oxford.

Zhao, F.J., McGrath, S.P., Merrington, G., 2007. Estimates of ambient background concentrations of trace metals in soils for risk assessment. Environ. Pollut. 148, 221-229.

Figure Captions

Figure 1 Spatial extent of soil and stream sediment sample PHE data from the G-BASE and Wolfson surveys, and locations of an independent validation set of topsoil As measurements used in this study across England and Wales.

Figure 2 Plots of geometric mean (GM) topsoil (<2 mm) versus GM deeper soil (<150 μm) for collocated samples grouped by PM group across area A+B (Figure 1) and linear regressions (---) for: a) As, b) Cr, and c) Ni.

Figure 3 Summary of approach to the estimation of soil equivalent PHE concentrations based on stream-sediment PHE concentrations for areas C and D shown in Figure 1. Calculations based on existing data in *italics;* transformations of data based on statistical relationships in **bold.**

Figure 4 Example of bedrock geology and superficial deposits separated into unique parent material (PM) combinations and their codes, and separate polygons within 1 kilometre squares of the British National Grid in Northamptonshire (UK). Individual polygon centroids are the centres for each 1-km PM polygon; the average centroid is the centre of the four individual polygon centroids shown.

Figure 5 Illustration showing the complexity of the As-rich Marlstone Rock Formation PM outcrop in central England with stream sediment and soil sample locations in a 10 kilometre square of the British National Grid.

Figure 6 Scatterplots for: i) percentiles (•) of the As distribution for Wolfson and G-BASE stream sediment data (Areas A+B+C), and ii) GM As concentrations grouped by PM (•) for which n>8. The least squares linear regression (---) was fit to the set of paired percentiles.

Figure 7 Scatterplot of MSD (mean squared differences) from 10 % holdout validations (HV) for estimated and actual GM As concentrations for samples with the same PM code based on the mean of *n* nearest neighbouring samples for: a) two repeated HV for soil samples

developed over PM (see Figure 2) with elevated As concentrations, b) three repeated HV using subsets of stream sediment samples from the Silurian Llandovery mudstone-dominated sedimentary rocks of Central Wales (no Quaternary deposits) which exhibit considerable lateral variation in stream sediment As concentrations.

Figure 8 Scatterplots of GM As in stream sediments versus GM As in soil plotted on a logarithmic scale for different PM groupings, and their second order polynomial regression models (---) for: a) 1km-PM polygons (n=29416), b) 1km-PM polygons averaged over 5-km grid squares (n=4025), and c) PM with data grouped by 5, 10km, 100km grid square, and geological map sheet (n=1188).

Figure 9 Categorical map of mineral topsoil equivalent geometric mean As ABCs (mg kg^{-1}) for England and Wales.

Figure 10 – Least squares non-linear regression model for the relationship between topsoil As GMs for individual PM groups and the estimated proportion of samples exceeding the regulatory threshold based on a log-normal distribution.

Table 1 Summary of the soil and stream sediment geochemical survey data used in this study (with reference to areas shown in Figure 1: G-BASE topsoil (A) GBASE deeper soil (A+B), GBASE sediments (A+B+C) and Wolfson Atlas (A+B+C+D).

Survey	Area in Figure 1	Number of samples	Mean sampling intensity	Soil sample depth (cm)	Size fraction analysed	Elements determined (including As, Cr, Ni)	[c] Analytical method -As	[c] Analytical method – Cr & Ni	Survey dates
[a] GBASE topsoil	A	6332	1 per 2 km^2	0-15	<2mm	45 major and trace	XRFS	XRFS	1994- 1996
[a] GBASE deeper soil	A+B	20,302	1 per 2 km^2	35-50	<150µm	45 major and trace	XRFS	XRFS (DR-ES[d])	1988 – 2000
[a] GBASE sediments	A+B+C	43,088	1 per 1.5 km^2	n/a	<150µm	Between 30 and 45 major and trace	XRFS (AAS[d])	XRFS(DR-ES[d])	1977- 2000
[b] Wolfson sediments	A+B+C+D	50,000	1 per 2.5 km^2	n/a	<177µm	19 major and trace	$KHSO_4$ fusion; Gutzeit method	DR-ES	1969

[a] Johnson et al., 2005; [b]Webb *et al.*, 1978

[c] XRFS (X-ray Fluorescence Spectrometry); DR-ES (Direct Reading Emission Spectrometry); AAS (Atomic Absorption Spectrometry: ammonium persulphate and 75% HCl acid digestion and solvent extraction); [d] samples collected north of Area A in Figure 1.

Table 2 Coefficients for regression equations relating collocated deeper soil and topsoil GM PHE concentrations grouped by PM across area A+B for: a) As, b) Cr, c) Ni (shown in Figure 2)

Dependent variable (y)	Independent variable (x)	Intercept (α) ± Std. error	Coefficient (β) ± Std. error	R^2	Number of observations	Standard error of observation
a) GM Topsoil As for PM group	GM deeper soil As for PM group	0.24 (0.56)	1.01 (0.03)	0.89	186	4.6
b) GM Topsoil Cr for PM group	GM deeper soil Cr for PM group	-1.85 (3.72)	0.88 (0.04)	0.72	176	13.8
c) GM Topsoil Ni for PM group	GM deeper soil Cr for PM group	-2.92 (1.17)	0.86 (0.04)	0.76	176	5.7

Table 3 Summary statistics and proportion of variance (%) accounted for in log transformed PHE concentrations in soils and stream sediments for areas shown in Figure 1.

	As	Cr	Ni
Topsoil (<2mm, n=6332) Area A			
Min.	1	1	1
Mean	16	74	23.5
Geometric mean	13.6	67	19.8
Geometric SD	1.68	1.6	1.9
Max.	342	2534	459
Skewness	8.4	28.3	6.1
Log$_e$ transformed skewness	0.62	-0.68	-0.79
[a]Variance (%) accounted for by PM classification	34.7	30.1	42.9
Deeper soil (<150 µm; n=20,302) Area A+B			
Min.	0.45	2	0.5
Mean	17.6	95	34.2
Geometric mean	14.1	88.4	29.9
Geometric SD	1.8	1.4	1.7
Max.	463.8	6787	7804
Skewness	8.7	48.8	117
Log$_e$ transformed skewness	0.67	-0.55	-0.51
[a]Variance (%) accounted for by PM classification	39.2	28.5	27
GBASE sediments (<150 µm; n=10,322) Area A+B			
Min.	1	1	1
Mean	17.9	97	42.7
Geometric mean	14.8	92.4	39.1
Geometric SD	1.8	1.4	1.5
Max.	407	2144	1789
Skewness	7.6	17.9	27.2
Log$_e$ transformed skewness	-0.1	-1.2	-0.2
[a]Variance (%) accounted for by PM classification	24.2	25.1	20

[a] using ANOVA applied to the log$_e$ transformed data

Table 4 Summary statistics (mg kg^{-1}) for G-BASE (GB) and Wolfson (WS) sediments for areas A+B+C in Figure 1

	GB As	WS As	GB Cr	WS Cr	GB Ni	WS Ni
Min.	1	1	1	1	1	1
Mean	25.44	17.2	107	58.8	38.9	32.1
Median	13.0	12.0	97.9	51.0	36.1	28.0
Geometric mean	15.0	10.9	99.6	43.8	34.5	24.8
Geometric SD	2.37	2.25	1.41	2.31	1.63	2.16
Max.	12400	4000	14590	30810	32390	3275
Skewness	82.5	42.35	93.9	150	13.7	37.2
Log$_e$ transformed skewness	0.82	0.75	-0.05	-1.46	-0.57	-0.87
n	27387	31006	32328	31006	32329	31006

Table 5 Coefficients for regression equations between GM_{sed} and GM_{soil} based on more than 4 samples in each PM group at different scales: a) nearest 5 samples for 1-km polygons, b) grouped by 5 km grid square, and c) grouped by nested-scale approach (see text). The regression plots are shown in Figures 8a-c.

Dependent variable (y)	Independent variable (x)	Intercept (α) (± Std. Error)	Coefficient β_1 (± Std. Error)	Coefficient β_2 (± Std. Error)	R^2	Number of observations	Standard error of observation
a) GM As deeper soil (nearest 5) by PM class	GM As sediment (nearest 5) by PM class	1.03 (0.04)	0.51 (0.03)	0.03 (0.01)	0.42	29416	0.31
b) GM As deeper soil (5 km grid square) by PM class	GM As sediment (5 km grid square) by PM class	1.60 (0.11)	0.03 (0.08)	0.13 (0.01)	0.47	4025	0.31
c) GM As deeper soil (nested-scales) by PM class	GM As sediment (nested-scales) by PM class	1.31 (0.20)	0.17 (0.15)	0.12 (0.03)	0.55	1188	0.27

Table 6 Mean Square Deviation (MSD) and bias statistics for GM_{soil} and soil equivalent $GMsoil_{est}$ for i) (areas A+B) after application of regression equations (Figures 8 a-c) to 1-km PM polygons based on three different scales of combining stream sediment data by PM polygon for estimation of soil equivalent As ABCs, and ii) independent validation data.

		Mean Squared Deviation ($\log_e$ concentration)			
Soil As concentration Range (mg kg^{-1})	Number of 1km-PM centroids	1-km polygons (cf. Fig 8a)	1km-polygons avg. by 5km^2 (cf Fig 8b)	Nested-scale approach (cf. Fig 8c)	Independent validation data (n=41)
< 10	41596	*0.214	* 0.214	0.237	n/a
10-20	111917	0.079	* 0.072	0.095	0.248
20-30	21734	0.196	0.194	*0.186	0.252
30-40	5662	0.486	0.502	*0.460	0.442
40-60	3489	0.683	0.761	*0.624	0.936
60-90	1405	1.122	1.266	*1.012	0.502
>=90	1104	2.016	2.366	*1.793	2.080
Bias		0.061	0.058	*0.032	0.191

* minimum MSD and bias for As concentration range

Table 7 Mean Square Deviation (MSD) for measured soil As (GMsoil) and estimated GM soil (GMsoil$_{est}$) based on sediment data after application of the regression equation from the nested approach (Fig 8c) for 1km-PM polygon centroids using i) GM, and ii) inverse distance weighted interpolation based on natural log transformed data.

Distance to furthest sample – range	Count		IDW
(km)	soil samples	GM	interpolation
		MSD	MSD
0 - 2.5	3129	**0.19	0.21
2.5 – 5	7377	**0.22	0.24
5 – 10	5846	**0.27	0.29
10 – 20	4163	**0.32	0.34
20 – 40	1370	**0.33	0.37
*40 – 80	663	**0.27	0.29
*80 – 160	170	**0.40	0.42
*>160	29	**0.46	0.56

* sediment data not grouped by PM at distances greater than 50 km.

** smallest MSD for distance class

Table 8 Geometric mean As ABC and 95% confidence intervals (±) for concentration ranges shown in Figure 9 for areas where GM As has been estimated using soil or sediment values (Soil$_{est}$). All units are mg kg^{-1}.

	Soil		Soil$_{est}$	
GM As Concentration range	GM As for concentration range	95% confidence interval (±)	GM As for concentration range	95% confidence interval (±)
<15	12	4	10	5
15-20	17	6	17	9
20-30	23	9	24	17
>30	48	38	59	62

Figure 1

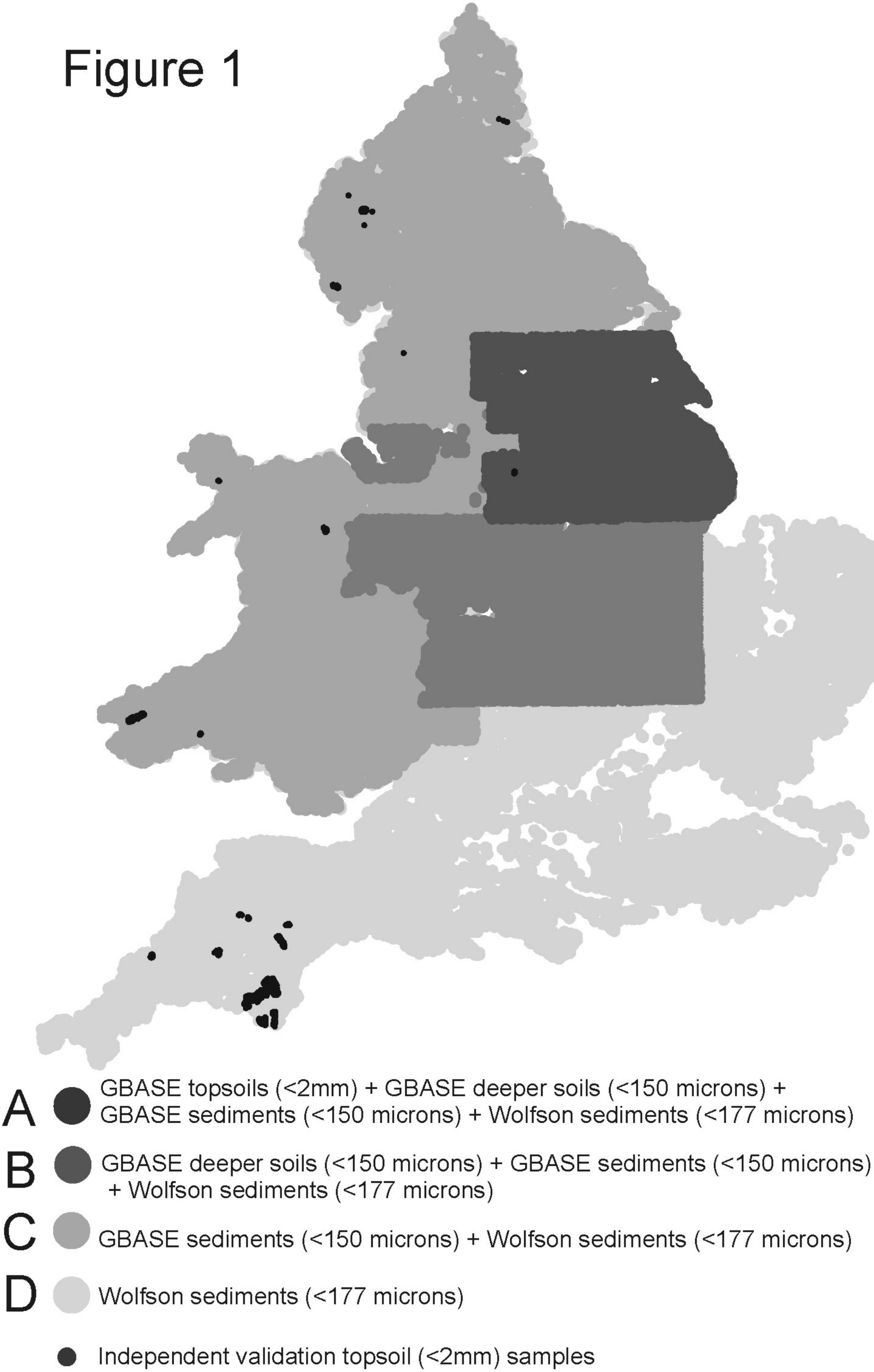
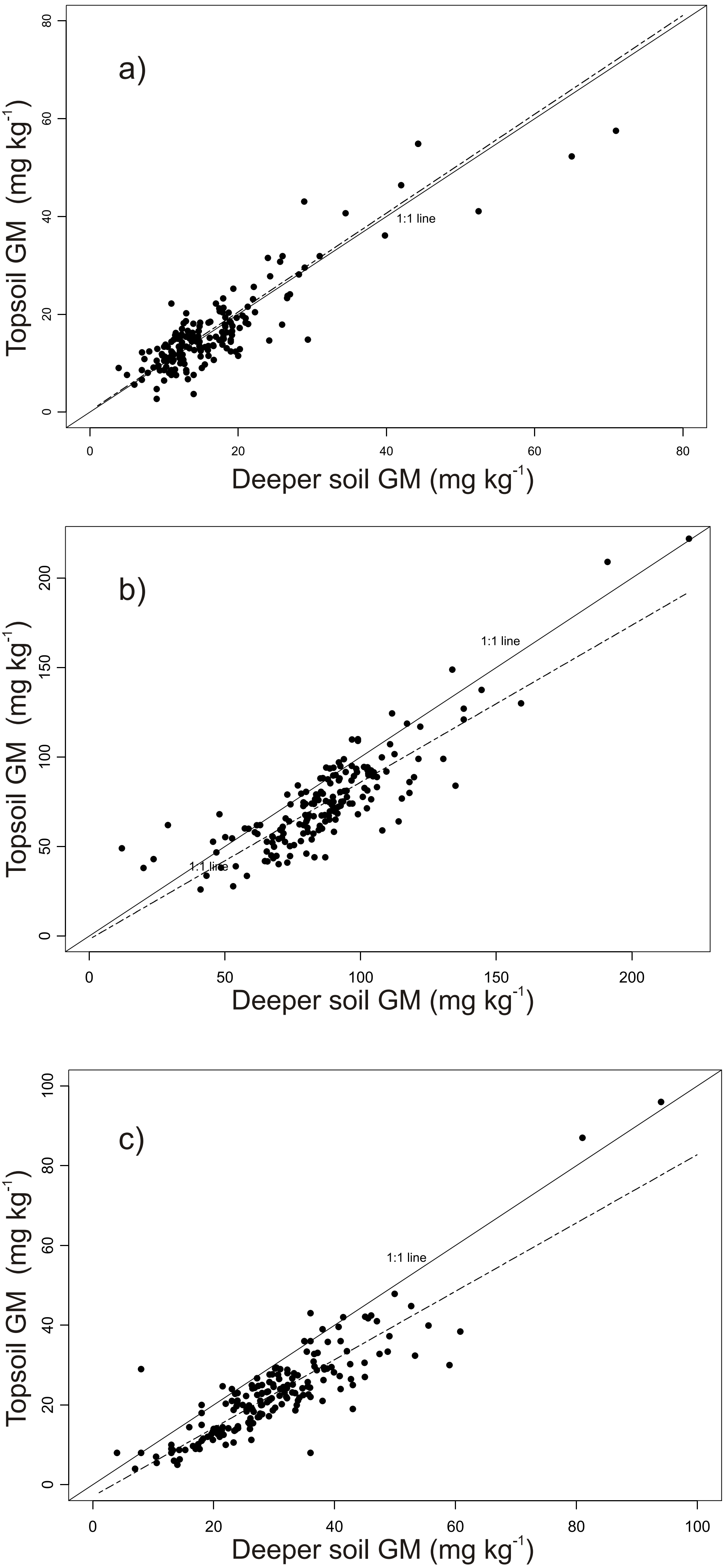
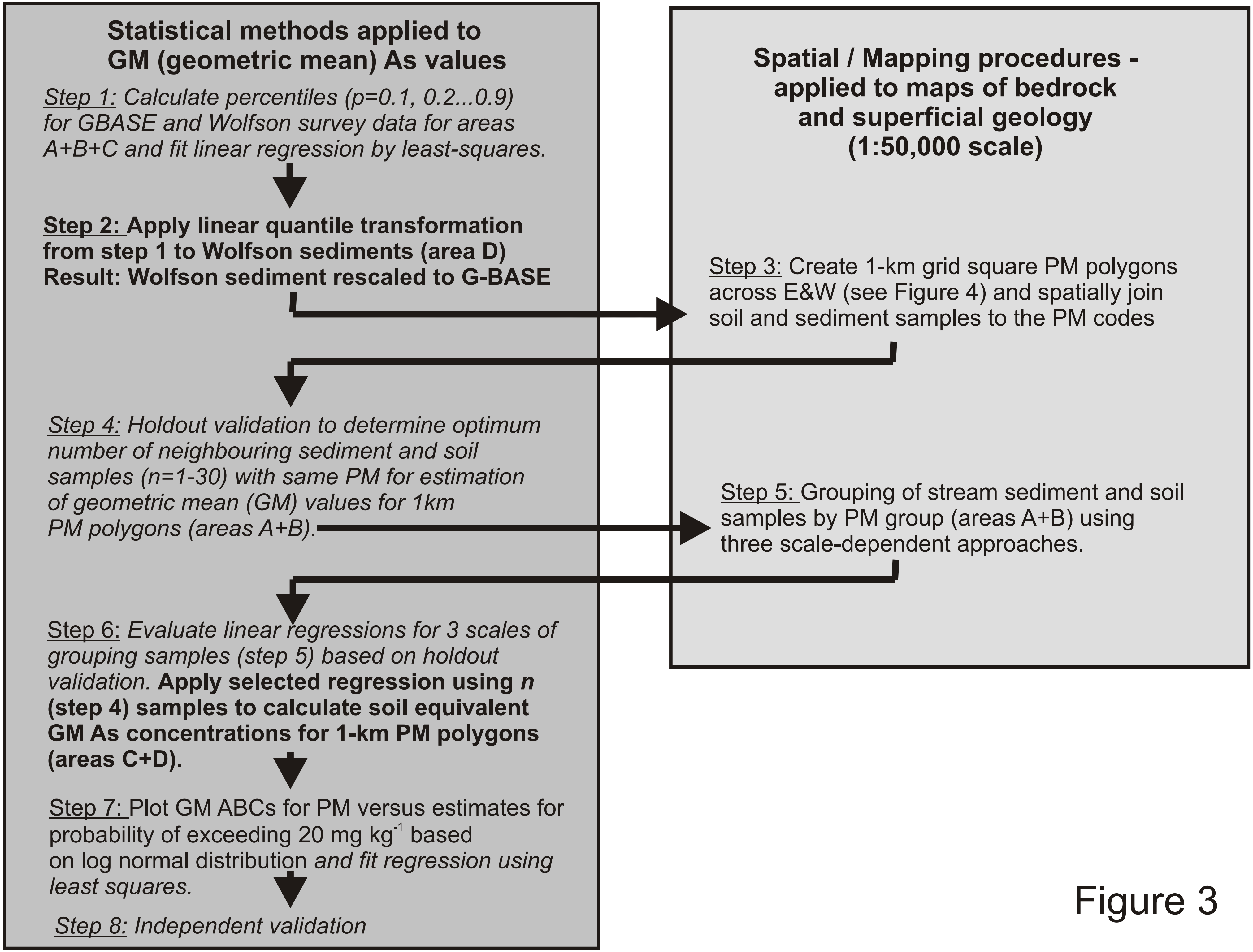

A ● GBASE topsoils (<2mm) + GBASE deeper soils (<150 microns) + GBASE sediments (<150 microns) + Wolfson sediments (<177 microns)
B ● GBASE deeper soils (<150 microns) + GBASE sediments (<150 microns) + Wolfson sediments (<177 microns)
C ● GBASE sediments (<150 microns) + Wolfson sediments (<177 microns)
D ● Wolfson sediments (<177 microns)
● Independent validation topsoil (<2mm) samples

Figure 2


a)
1:1 line
Topsoil GM (mg kg^{-1})
Deeper soil GM (mg kg^{-1})
0
20
40
60
80
0
20
40
60
80
b)
1:1 line
1:1 line
Topsoil GM (mg kg^{-1})
Deeper soil GM (mg kg^{-1})
0
50
100
150
200
0
50
100
150
200
c)
1:1 line
Topsoil GM (mg kg^{-1})
Deeper soil GM (mg kg^{-1})
0
20
40
60
80
100
0
20
40
60
80
100



Statistical methods applied to GM (geometric mean) As values

Step 1: *Calculate percentiles (p=0.1, 0.2...0.9) for GBASE and Wolfson survey data for areas A+B+C and fit linear regression by least-squares.*

Step 2: Apply linear quantile transformation from step 1 to Wolfson sediments (area D) Result: Wolfson sediment rescaled to G-BASE

Step 4: *Holdout validation to determine optimum number of neighbouring sediment and soil samples (n=1-30) with same PM for estimation of geometric mean (GM) values for 1km PM polygons (areas A+B).*

Step 6: *Evaluate linear regressions for 3 scales of grouping samples (step 5) based on holdout validation.* **Apply selected regression using *n* (step 4) samples to calculate soil equivalent GM As concentrations for 1-km PM polygons (areas C+D).**

Step 7: Plot GM ABCs for PM versus estimates for probability of exceeding 20 mg kg^{-1} based on log normal distribution *and fit regression using least squares.*

Step 8: *Independent validation*

Spatial / Mapping procedures - applied to maps of bedrock and superficial geology (1:50,000 scale)

Step 3: Create 1-km grid square PM polygons across E&W (see Figure 4) and spatially join soil and sediment samples to the PM codes

Step 5: Grouping of stream sediment and soil samples by PM group (areas A+B) using three scale-dependent approaches.

Figure 3

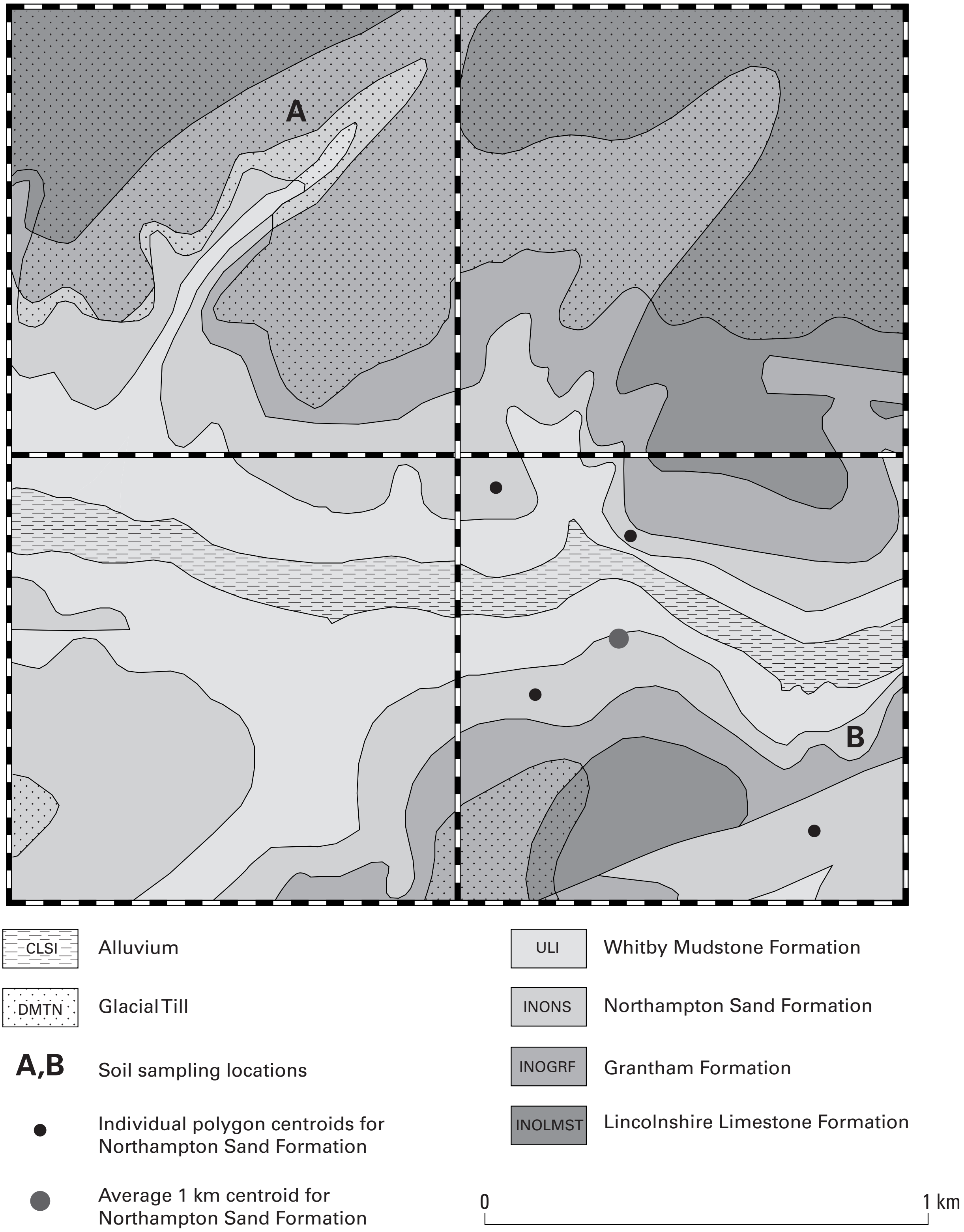

A
B
CLSI
Alluvium
DMTN
Glacial Till
A,B
Soil sampling locations
Individual polygon centroids for Northampton Sand Formation
Average 1 km centroid for Northampton Sand Formation
ULI
Whitby Mudstone Formation
INONS
Northampton Sand Formation
INOGRF
Grantham Formation
INOLMST
Lincolnshire Limestone Formation
0
1 km

Figure 4

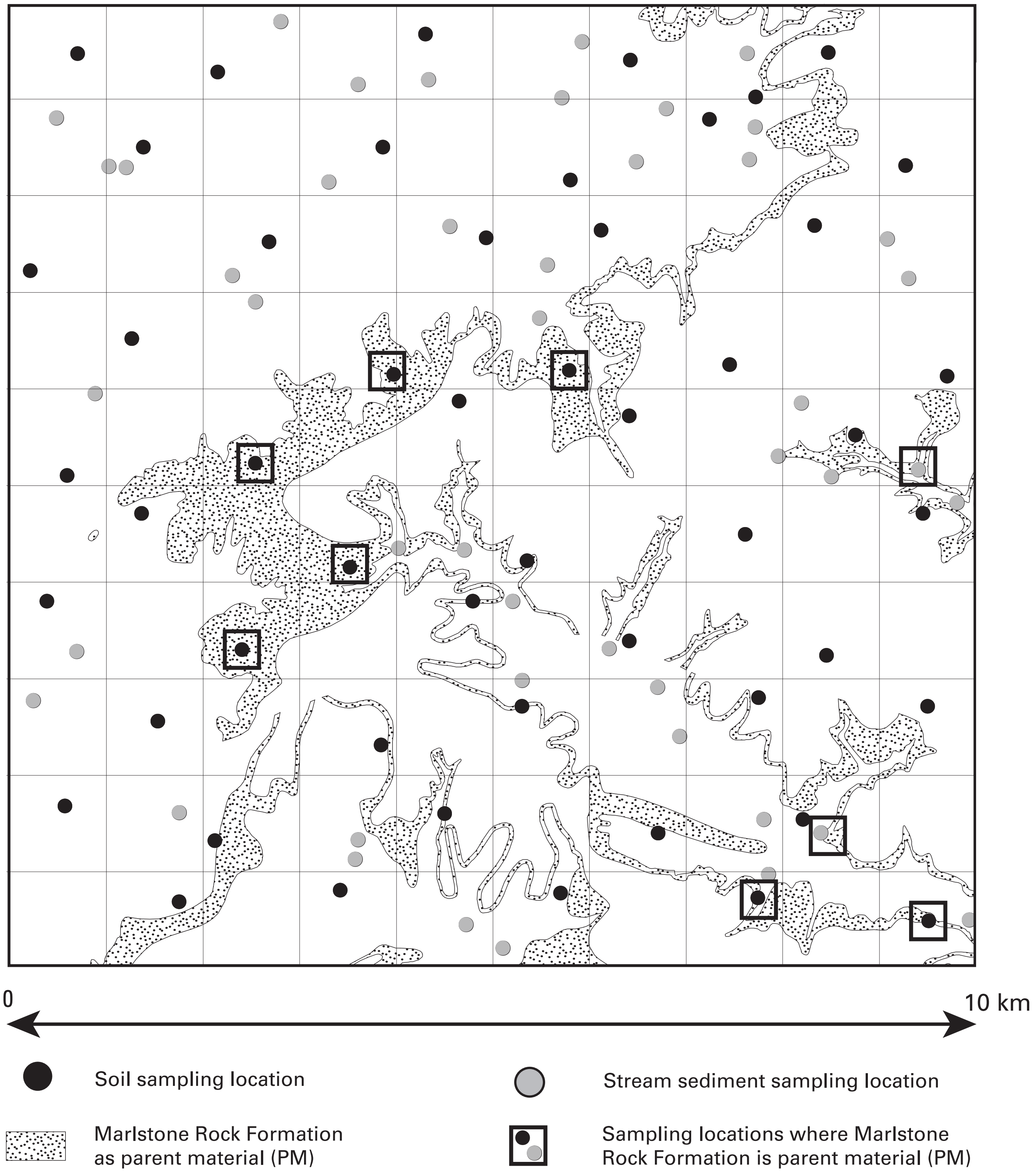
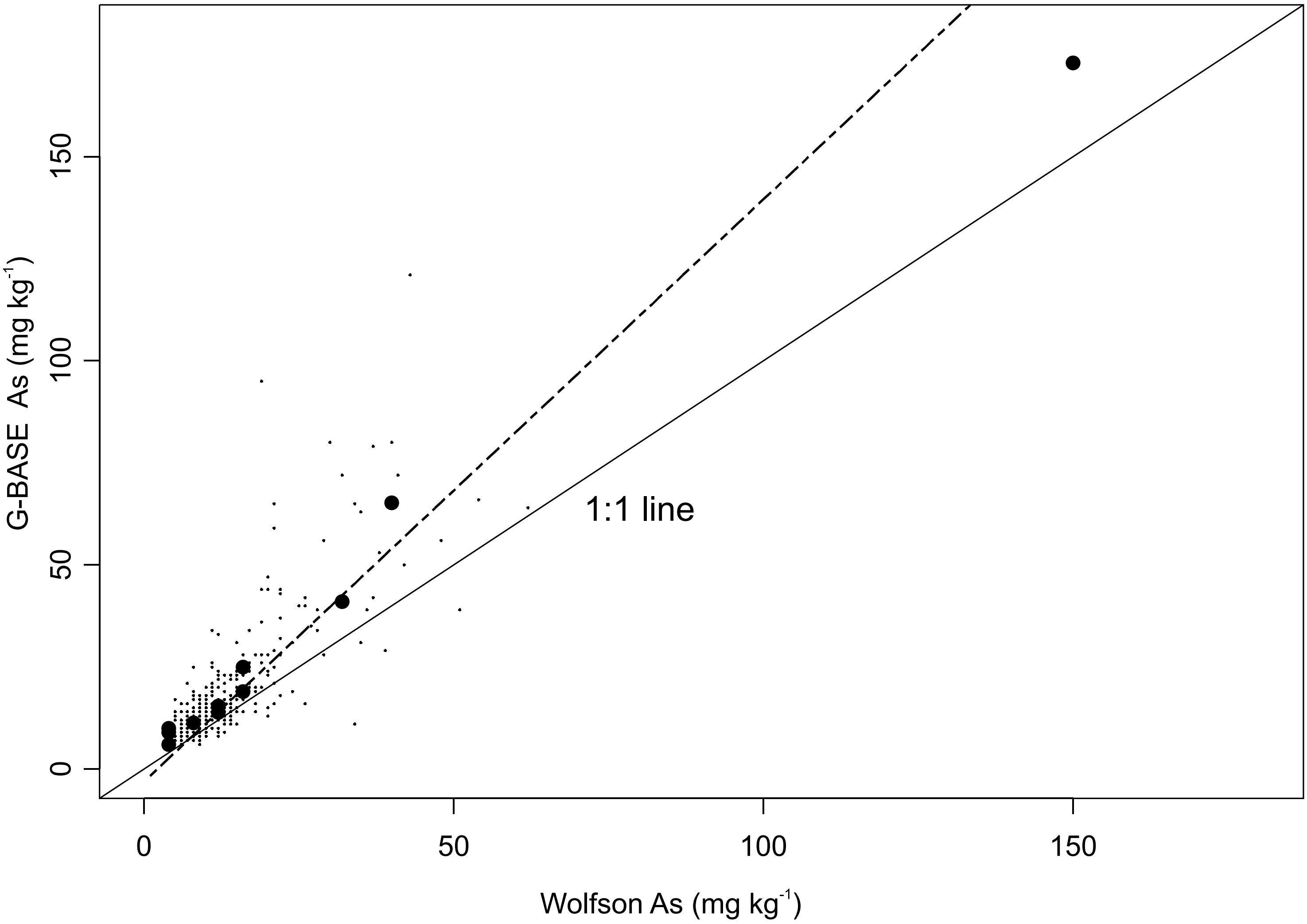

0
10 km
Soil sampling location
Stream sediment sampling location
Marlstone Rock Formation as parent material (PM)
Sampling locations where Marlstone Rock Formation is parent material (PM)

Figure 5

Figure 6



G-BASE As (mg kg^{-1})
Wolfson As (mg kg^{-1})
0
50
100
150
0
50
100
150
1:1 line

Figure 7

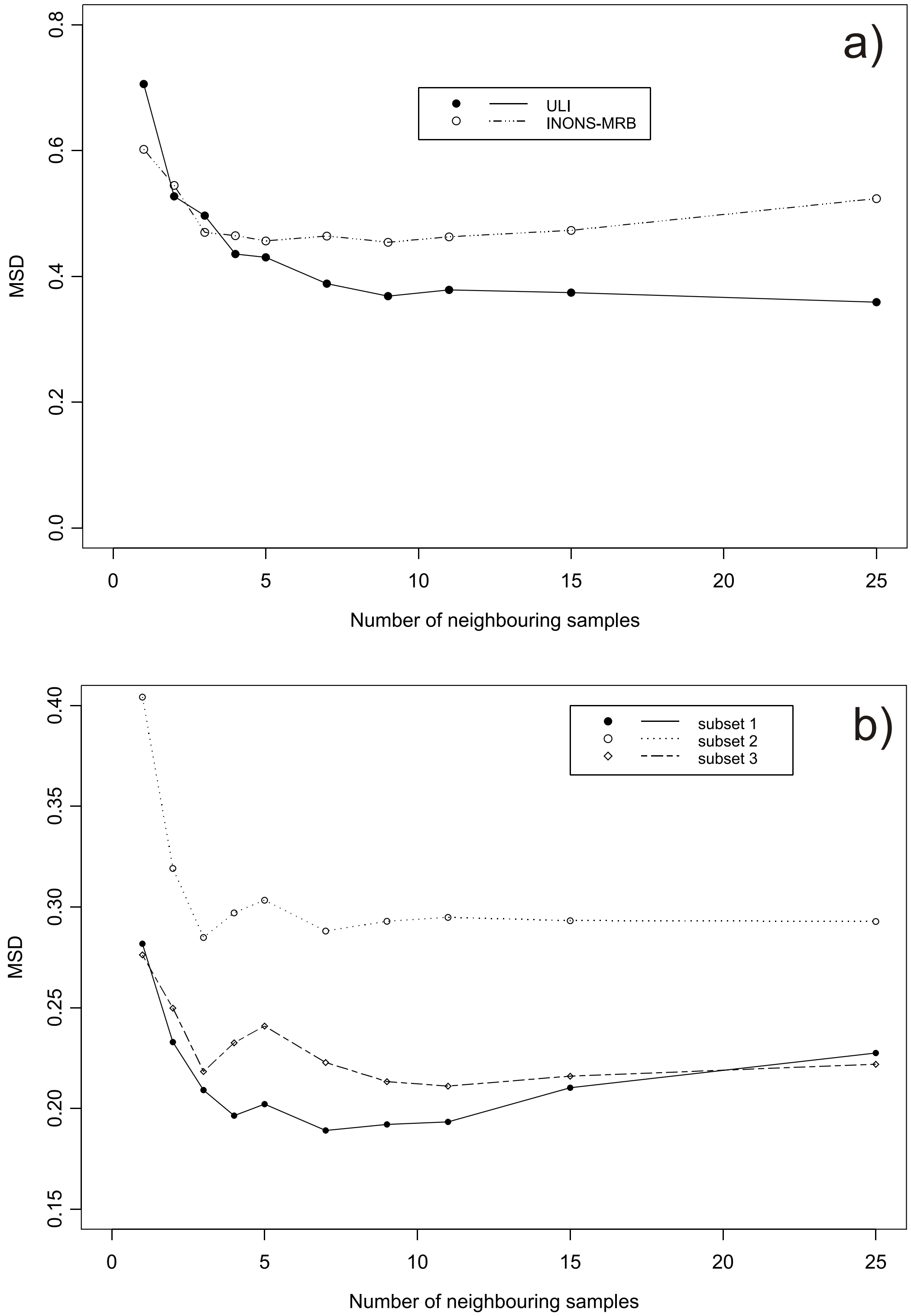
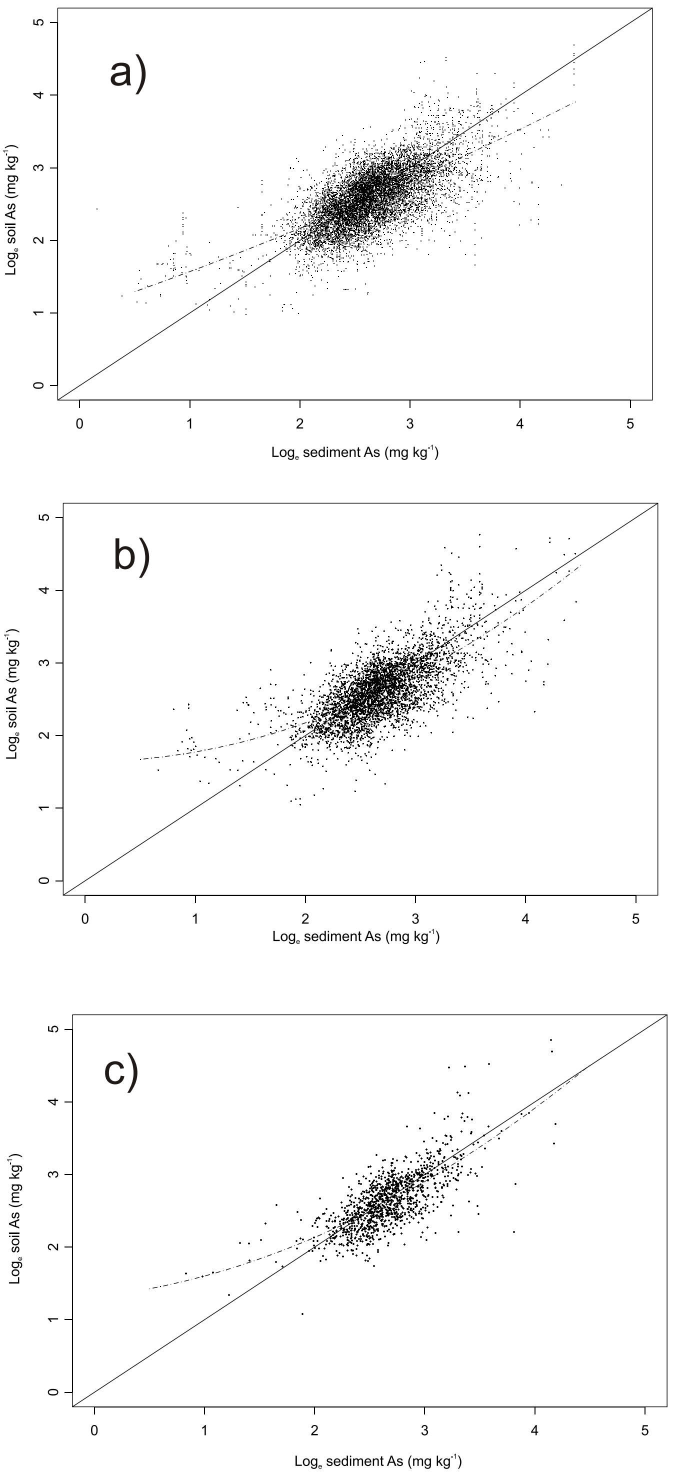

a)
ULI
INONS-MRB
MSD
0.0
0.2
0.4
0.6
0.8
0
5
10
15
20
25
Number of neighbouring samples
b)
subset 1
subset 2
subset 3
MSD
0.15
0.20
0.25
0.30
0.35
0.40
0
5
10
15
20
25
Number of neighbouring samples

Figure 8



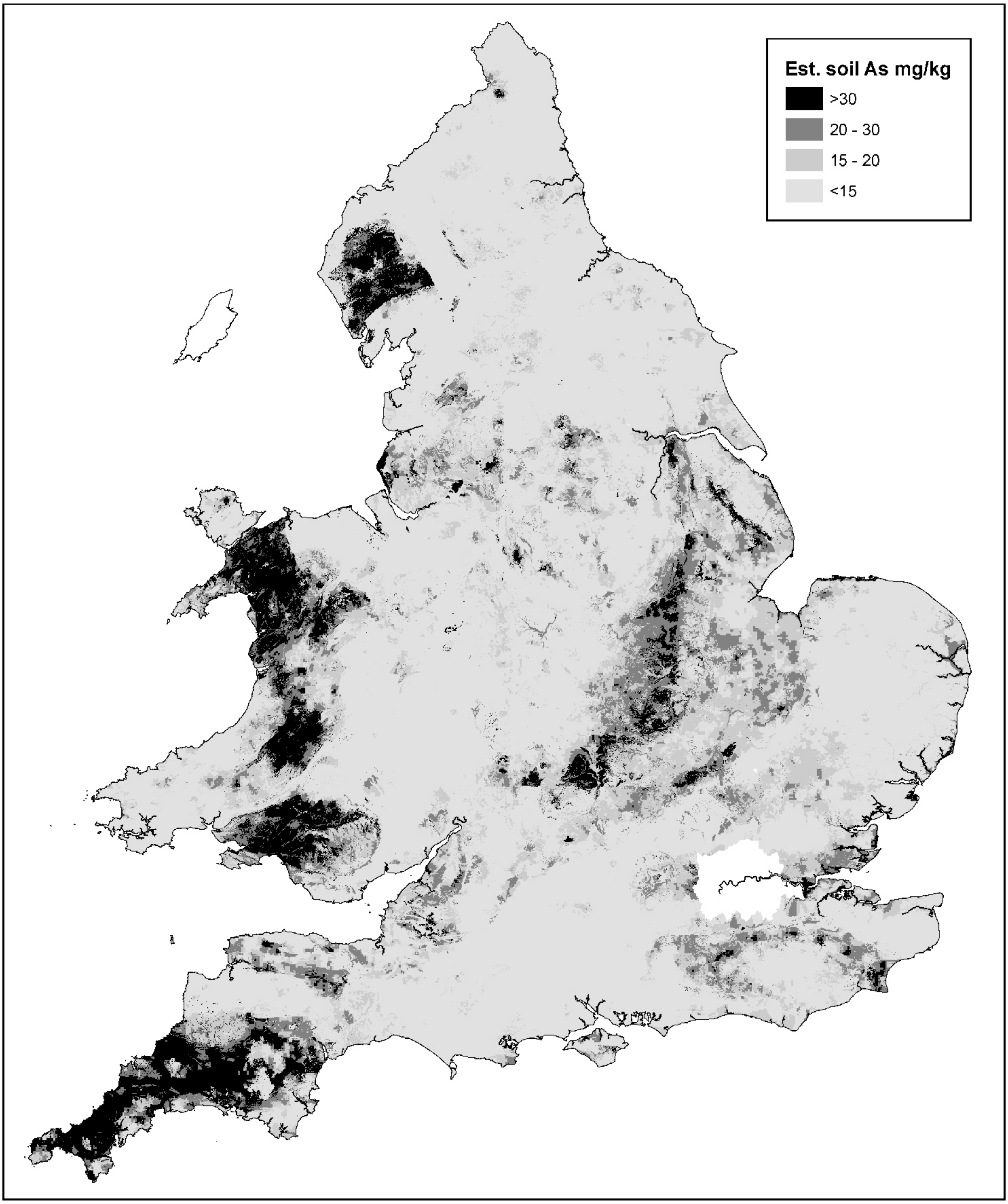
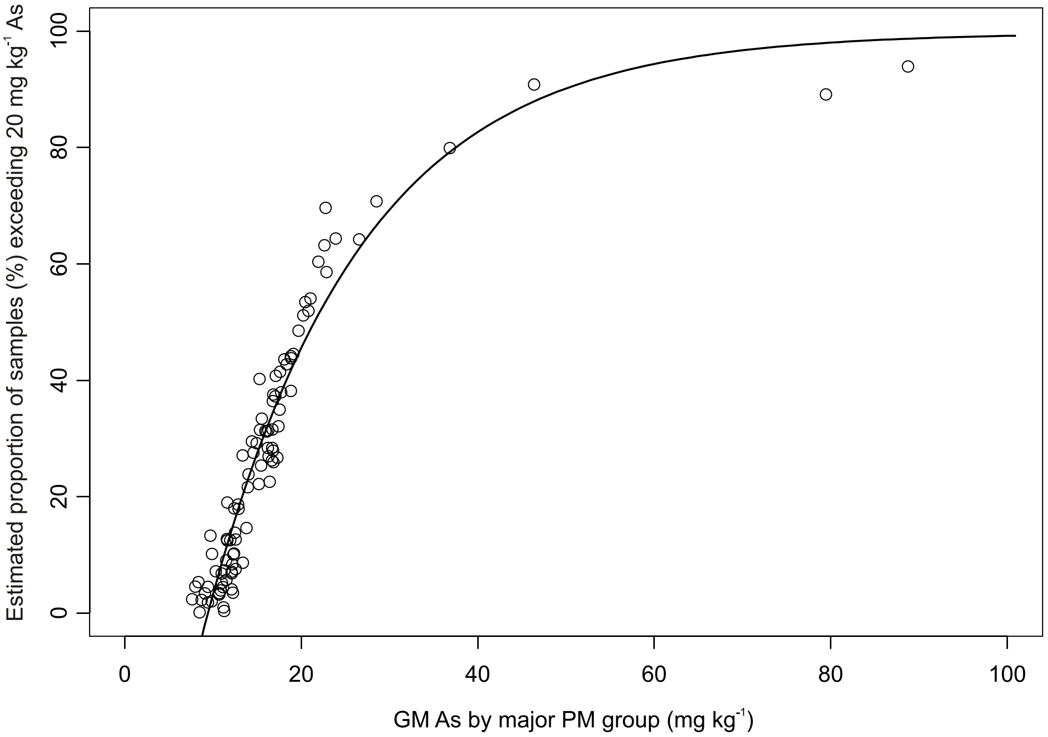


Figure 9

Figure 10


Estimated proportion of samples (%) exceeding 20 mg kg^{-1} As
0
20
40
60
80
100
GM As by major PM group (mg kg^{-1})