

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

B. P. MARCHANT<sup>a\*</sup>, A. M. TYE<sup>b</sup> & B. G. RAWLINS<sup>b</sup>

<sup>a</sup>*Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ* and <sup>b</sup>*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](mailto:ben.marchant@bbsrc.ac.uk)

## 1 **Summary**

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

## 20 **Introduction**

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

35 Generally, regional estimates of the contribution of natural sources to metal con-  
36 centrations in contaminated soil are made from the summary statistics of surveys made  
37 in areas which are assumed to be unaffected by anthropogenic processes. It is possi-  
38 ble to distinguish between natural and anthropogenic sources of some elements such  
39 as lead by the stable isotopes (Clark *et al.*, 2006) but in other cases the metals may  
40 only have one stable isotope or analytical methods may not be widely available for the  
41 determination of isotope fractions (e.g. copper and tin). Hamon *et al.* (2004) tested  
42 whether various soil properties could be used as indicators of the background or nat-  
43 ural metal content of contaminated soils. They found that the natural concentrations  
44 of arsenic, chromium, cobalt, copper, lead, nickel, and zinc could be approximated in  
45 terms of the iron and manganese concentrations in the soil. Their tests were conducted  
46 in south-east Asia but they suggest that these relationships may hold worldwide. This  
47 approach assumes that the iron content of contaminated soils is not elevated by an-

48 thropogenic processes. Such behaviour has been observed in previous surveys of urban  
49 soil contamination in the UK. For example, Figure 1 shows that metal processing in  
50 Sheffield has enriched the lead content of the soils in comparison with uncontaminated  
51 rural soils, but the iron content is relatively unchanged.

52 Conventional geostatistical methods are most efficient when the property being  
53 mapped approximates, or may be transformed to approximate, a Gaussian distribution.  
54 However point-sources of pollution lead to hotspots or outliers in the distribution of soil  
55 metals which are inconsistent with the Gaussian assumption. Therefore robust geosta-  
56 tistical methods have been applied to surveys of soil metal pollution. Robust methods  
57 estimate the statistics of the underlying variation of metal concentrations with mini-  
58 mum effect of outliers. In geostatistical analysis we first estimate a variogram model  
59 which describes the spatial variation of the property of interest based upon the obser-  
60 vations. This model is then used to predict the property at unsampled locations. In  
61 conventional geostatistics the variogram model is estimated by Matheron's method of  
62 moments estimator (Webster & Oliver, 2007). This estimator is sensitive to outlying  
63 observations. Therefore robust variogram estimators have been devised that model the  
64 underlying variation in the presence of outliers. Three such robust estimators were  
65 compared by Lark (2000). Lark (2002) suggested a statistic which may be used to  
66 identify outlying observations. This statistic was used to identify outliers in surveys of  
67 heavy metal contamination in Sheffield, UK (Rawlins *et al.*, 2005) and Zhangjiagang,  
68 China (Zhao *et al.*, 2007). The outliers were removed from the datasets before the  
69 diffuse pollution was predicted across these study regions. However, although outliers  
70 are likely to be dominated by point-source pollution they may still contain information  
71 about the diffuse pollution. Therefore Marchant *et al.* (2010) used a robust prediction  
72 algorithm (Hawkins & Cressie, 1984) to winsorize the observations. This winsorizing  
73 process separated each observation into two components, one because of localized pro-  
74 cesses and one because of diffuse processes. A similar approach was applied by Papritz  
75 (2007) when mapping pollution around a Swiss smelter.

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

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

## 101 **Theory**

### 102 *Geostatistical Prediction of Soil Properties*

103 The variation of a soil property may be described by the linear mixed model (LMM)

104 which divides the spatial variation between fixed and random effects (Lark & Cullis,  
 105 2004) and accounts for variation between observations made at the same site which  
 106 we may think of as measurement error. The fixed effects are a linear combination of  $q$   
 107 covariates and represent variation of the expectation of the property across the study  
 108 region. The random effects describe the spatially correlated component of variation of  
 109 the property. The LMM is written

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

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

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

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

127 For a second order stationary random variable

$$C(h) = C(0) - \gamma(h). \quad (3)$$

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

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

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

133 where  $c_0$  is the nugget variance,  $c_1$  is the partial sill variance,  $a$  is a distance parameter,  
 134  $\nu$  is a smoothness parameter,  $K_\nu$  is a modified Bessel function of the second kind of  
 135 order  $\nu$  (Abramowitz & Stegun, 1972) and  $\Gamma$  is the gamma function.

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

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

153 calculate

$$\theta_i = \frac{\{z(\mathbf{x}_i) - \tilde{Z}_{(-i)}\}^2}{\sigma_{(-i)}^2}, \quad (5)$$

154 where  $\tilde{Z}_{(-i)}$  and  $\sigma_{(-i)}^2$  denote the prediction and prediction variance at  $\mathbf{x}_i$  when  $z(\mathbf{x}_i)$   
155 is omitted from the transformed observation vector. If the fitted model is a valid  
156 representation of the spatial variation of the soil property and the prediction errors are  
157 Gaussian then  $\boldsymbol{\theta} = [\theta_1 \dots \theta_n]^T$  is a realization of a  $\chi_1^2$  distribution with mean  $\bar{\boldsymbol{\theta}} = 1.0$   
158 and median  $\check{\boldsymbol{\theta}} = 0.455$ . Quantile-quantile (QQ) plots of the  $(\theta_i)^{\frac{1}{2}}$  can be drawn to  
159 confirm that the assumption of Gaussian errors is reasonable.

### 160 *Robust Geostatistical Methods*

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

172 Robust method of moments variogram estimators have been devised by Cressie  
173 & Hawkins (1980), Dowd (1984) and Genton (1998). The methods make robust point  
174 estimates of the variogram of the underlying variation. Lark (2000) tested these esti-  
175 mators by looking at validation statistics of variogram models fitted to simulated data.  
176 He suggested that  $\check{\boldsymbol{\theta}}$  was a suitable robust statistic to assess the fitted variograms. Lark  
177 (2000) found that Matheron's estimator outperformed the robust estimators when the  
178 property was not contaminated. However when there was contamination each of the  
179 robust estimators outperformed Matheron's estimator. The relative performance of the



180 robust estimators varied with the form of contamination.

181 Lark (2002) suggested that once a robust variogram model has been fitted, out-  
 182 liers could be identified by a threshold on the  $\theta_i$  from cross-validation. Rawlins *et al.*  
 183 (2005) followed this approach and removed outliers before predicting soil metal con-  
 184 centrations at unsampled sites. However the removal of entire observations discards  
 185 information about the underlying process. Therefore, when analysing a survey of soil  
 186 metal contamination across France, Marchant *et al.* (2010) used a winsorizing algo-  
 187 rithm suggested by Hawkins & Cressie (1984) to divide each observation between a  
 188 component from underlying processes and a component from the secondary processes.  
 189 They then applied the BLUP to the underlying variation and mapped the observations  
 190 of the secondary process separately. The steps of this winsorizing algorithm are

- 191 1. Estimate a robust variogram of  $\mathbf{z}$ .
- 192 2. Compute the BLUP weights  $\lambda_{j(-i)}$ ,  $j = 1, \dots, i - 1, i + 1, \dots, n$  required for  
 193 leave-one-out cross validation and the corresponding kriging variance  $\sigma_{(-i)}^2$ .
- 194 3. Compute the weighted median  $\check{z}_{(-i)}$  for  $i = 1 \dots n$ . The weighted median solves  
 195  $\sum_{j=1, j \neq i}^n \lambda_{j(-i)} \text{sign} \{ \check{z}(\mathbf{x}_i) - z(\mathbf{x}_j) \} = 0$ , where  $\text{sign}(y) = -1$  for  $y < 0$  and  
 196  $\text{sign}(y) = 1$  otherwise. This equation may have more than one solution but  
 197 Hawkins & Cressie (1984) state that the number of solutions is always odd and  
 198 therefore a unique solution can be defined by the median of these solutions.

4. Winsorize the data by replacing  $z_i$  by

$$z_c(\mathbf{x}_i) = \begin{cases} \check{z}_{(-i)} + c\sigma_{(-i)} & \text{if } z(\mathbf{x}_i) - \check{z}_{(-i)} > c\sigma_{(-i)} \\ z(\mathbf{x}_i) & \text{if } |z(\mathbf{x}_i) - \check{z}_{(-i)}| \leq c\sigma_{(-i)} \\ \check{z}_{(-i)} - c\sigma_{(-i)} & \text{if } z(\mathbf{x}_i) - \check{z}_{(-i)} < -c\sigma_{(-i)} \end{cases} \quad (6)$$

199 where  $c$  is a constant  $1.5 < c < 3.0$ .

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

202 Marchant *et al.* (2010) repeated the above algorithm for different values of  $c$  and  
 203 calculated cross-validation  $\theta$  statistics for each  $\mathbf{z}_c$ . The use of a robust variogram

204 estimator in stage 1 ensured that for large  $c$ ,  $\check{\theta} \approx 0.455$  but in the presence of outliers  
205  $\bar{\theta} > 1.0$ . The value of  $\bar{\theta}$  decreased more rapidly than  $\check{\theta}$  as  $c$  was decreased and their  
206 final prediction of the underlying variation was based upon the  $\mathbf{z}_c$  for which  $\bar{\theta}$  was  
207 closest to 1.0. In the original formulation of the Hawkins & Cressie (1984) algorithm  
208 the mean of  $\mathbf{z}$  was assumed to be constant and the BLUP in Step 2 was equivalent  
209 to ordinary kriging. Papritz (2007) expanded the algorithm to include fixed effects.  
210 The fixed effect coefficients were estimated by a robust regression estimator and the  
211 winsorizing algorithm was applied to the residuals.

## 212 **Methods**

### 213 *The Study Area*

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

229 In late 18<sup>th</sup> and early 19<sup>th</sup> century Swansea there were many smelters process-  
230 ing copper, arsenic, lead, zinc, silver and tin. The height of the chimney stacks was  
231 increased in the 19<sup>th</sup> century to disperse the toxic fumes from the copper smelters.

232 The lead-smelting industry was particularly significant in the 17<sup>th</sup> and 19<sup>th</sup> centuries,  
233 although compared to copper a greater proportion of smelting was undertaken in the  
234 ore fields. A total of 250 000 tonnes of raw copper-ore was processed in the Swansea  
235 Valley annually in the mid 19<sup>th</sup> century yielding 22 000 tonnes of refined copper; the  
236 dominant source of ore was Devon and Cornwall (Hughes, 2000). The copper industry  
237 was considered to be the principal contributor to Swansea's pollution problems. Newell  
238 & Watts (1996) used a Gaussian plume model to estimate annual average suspended  
239 airborne concentrations of arsenic, lead and tin during the mid-19<sup>th</sup> century in the  
240 vicinity of the Llanelli copper company 12 miles north-west of Swansea. The estimates  
241 were between 10 and 15  $\mu\text{g m}^{-3}$ . By contrast, current EC regulations stipulate limits of  
242 2  $\mu\text{g m}^{-3}$ . More recently remediation has been undertaken; the Lower Swansea Valley  
243 project of the 1960s and 1970s reclaimed slag heaps and large tracts of derelict land.

#### 244 *The Urban Survey*

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

254 Samples were collected according to the protocols of the Geochemical Surveys of  
255 Urban Environments (GSUE) project (Fordyce *et al.*, 2005) across Swansea, Neath,  
256 Port Talbot and the Mumbles area of the Gower Peninsula. Sample sites were selected  
257 from open ground as close as possible to the centre of each of four 500-metre squares,  
258 within each kilometre square of the British National Grid (BNG). Typical locations  
259 for sampling were gardens, parks, sports fields, road verges, allotments, open spaces,

260 schoolyards and waste ground. Each composite sample was based on nine samples  
261 of equal size from the corners, sides and centre of square of side-length 2 m. Each  
262 sample was collected at a depth range of 0-15 cm from the soil surface using an auger  
263 of diameter 35 mm. At each site, information was recorded on location using 1:10 000  
264 scale Ordnance Survey maps, a description of any visible contamination (e.g. metallic,  
265 pottery, bricks, plastics etc.), Munsell colour, soil clast lithologies (e.g. sandstone,  
266 limestone, etc.) and land use. All soil samples were disaggregated following air-drying  
267 and sieved to less than 2 mm. All samples were coned and quartered, and a 50-g  
268 subsample was ground in an agate planetary ball mill. The total concentrations of  
269 18 major and trace elements were determined by wavelength and energy dispersive  
270 X-ray fluorescence spectrometry (XRF-S). In this paper we only consider five elements  
271 (detection limits in parentheses): arsenic (1 mg kg<sup>-1</sup>), copper (1 mg kg<sup>-1</sup>), total iron  
272 expressed as Fe<sub>2</sub>O<sub>3</sub> (0.01 %), lead (2 mg kg<sup>-1</sup>), and tin (1 mg kg<sup>-1</sup>). For brevity we  
273 refer to these variables as metal concentrations although arsenic is a metalloid. Brief  
274 descriptions of the local land use at and around each site were tabulated for the years  
275 1900 and 2007 from Ordnance Survey maps of the area.

### 276 *The Rural survey*

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

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

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

### 297 *Statistical Analysis of Soil Metal Concentrations Around Swansea*

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

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

316 (1984) and Genton (1998). Cross-validation was performed for each fitted variogram  
317 and the estimator with  $\check{\theta}$  closest to 0.455 was selected. The observations were then  
318 winsorized according to the algorithm of Hawkins & Cressie (1984) for various values  
319 of constant  $1.5 < c < 3.0$ . This algorithm removes both positive and negative outliers.  
320 However we expect that the majority of outliers will be positive and caused by point  
321 source pollution. Therefore we only remove these positive outliers.

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

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

## 336 **Results**

### 337 *Prediction of Natural Metal Concentrations*

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

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

### 347 *Geostatistical Prediction of Anthropogenic Metal Concentrations*

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

### 370 *Distribution and magnitude of point and diffuse metal pollution*

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

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

## 389 Discussion

390 The survey confirms that the soils around Swansea remain substantially contaminated  
391 by historic metal and metalloid pollution. The soil metal concentrations cannot be  
392 represented by conventional geostatistical methods because the combination of diffuse  
393 and point-source pollution leads to complex patterns of variation. When conventional  
394 models were fitted to the data they were found to be invalid. The estimated variances  
395 were inflated by a small number of large observations at former industrial sites and  
396 thus it was not possible to accurately quantify the uncertainty of the predictions which  
397 result. However, plausible models did result when the diffuse and point-source pollution  
398 were mapped separately by robust geostatistical methods. In a previous survey, robust



399 methods were also required to map diffuse metal pollution around Sheffield (Rawlins  
400 *et al.*, 2005) and it is likely that that similar methods will be required to assess metal  
401 contamination in other industrial regions.

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

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

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

426       The pattern of sites where point-source pollution was identified is consistent with

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

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

449         In England and Wales the first tier of a human health or ecological risk assess-  
450 ment is a comparison between observed total soil metal concentrations at a site and  
451 their guideline values (Environment Agency, 2009) or screening values (Environment  
452 Agency, 2008). In the case of human health risk assessment, the revised Soil Guideline  
453 Values for arsenic concentrations in topsoil ( $32 \text{ mg kg}^{-1}$  for residential land use) is  
454 exceeded by the predicted sum of natural content and diffuse pollution for 89% of the

455 study area. Ecological health risks are assessed according to the difference between  
456 observed concentrations and ambient background metal concentrations (ABC) in soil.  
457 The proposed screening values for lead ( $167 \text{ mg kg}^{-1}$ ) and copper ( $88 \text{ mg kg}^{-1}$ ) are  
458 exceeded by the predictions of diffuse pollution for 44% and 58% of the study area  
459 respectively. When the ABCs are established it is important to ensure that they do  
460 not include any diffuse metal pollution.

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

## 471 **Conclusions**

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

482 **Acknowledgements**

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

490 **References**

491

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

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

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

501 Clark, H.F., Brabander, D.J. & Erdil, R.M. 2006. Sources, sinks, and exposure path-  
502 ways of lead in urban garden soil. *Journal of Environmental Quality*, **35**, 2066–  
503 2074.

504 Commission of the European Communities, 2006. Thematic Strategy for Soil Protec-  
505 tion. Brussels. [http://ec.europa.eu/environment/soil/pdf/com\\_2006\\_0231\\_en.pdf](http://ec.europa.eu/environment/soil/pdf/com_2006_0231_en.pdf).  
506 Accessed 25th March 2009.

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

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

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

- 515 Environment Agency, 2009. Soil screening values for assessing ecological risks. Ac-  
516 cessed 17th July, 2009. [http://www.environment-agency.gov.uk/static/documents/](http://www.environment-agency.gov.uk/static/documents/Research/ssv_2149429.pdf)  
517 [Research/ssv\\_2149429.pdf](http://www.environment-agency.gov.uk/static/documents/Research/ssv_2149429.pdf)
- 518 Fordyce, F.M., Brown, S.E., Ander, E.L., Rawlins, B.G., O'Donnell, K.E., Lister,  
519 T.R. Breward, N. & Johnson, C.C. 2005. GSUE: urban geochemical mapping in  
520 Great Britain. *Geochemistry: Exploration, Environment, Analysis*, **5**, 325–336.
- 521 Genton, M.G. 1998. Highly robust variogram estimation. *Mathematical Geology*, **30**,  
522 213–221.
- 523 Hamon, R.E., McLaughlin, M.J., Gilkes, R.J., Rate, A.W. Zarcinas, B., Robertson,  
524 A., Cozens, G., Radford, N. & Bettenay, L. 2004. Geochemical indices allow esti-  
525 mation of heavy metal background concentrations in soils. *Global biogeochemical*  
526 *cycles* **18**, GB 1014.
- 527 Hawkins, D. M. & Cressie, N. 1984. Robust kriging — A proposal. *Mathematical*  
528 *Geology*, **16**, 3–18.
- 529 Hughes, S. 2000. *Copperopolis – landscapes of the early industrial period in Swansea*.  
530 Royal Commission on the ancient and historical monuments of Wales. Cambrian  
531 Printers Limited, Ceredigion. pp. 358.
- 532 Lark, R.M. 2000. A comparison of some robust estimators of the variogram for use  
533 in soil survey. *European Journal of Soil Science*, **51**, 137–157.
- 534 Lark, R.M. 2002. Modelling complex soil properties as contaminated regionalized  
535 variables. *Geoderma*, **106**, 173–190.
- 536 Lark, R.M. & Cullis, B.R. 2004. Model based analysis using REML for inference  
537 from systematically sampled data on soil. *European Journal of Soil Science*, **55**,  
538 799–813.

- 539 Marchant, B.P. & Lark, R.M. 2007a. The Matérn variogram model: Implications for  
540 uncertainty propagation and sampling in geostatistical surveys. *Geoderma*, **140**,  
541 337–345.
- 542 Marchant, B.P. & Lark, R.M. 2007b. Optimal sampling for geostatistical surveys.  
543 *Mathematical Geology*, **39**, 113–134.
- 544 Marchant, B.P., Newman, S., Corstanje, R., Reddy, K.R., Osborne, T.Z. & Lark,  
545 R.M. 2009. Spatial monitoring of a non-stationary soil property: phosphorus in  
546 a Florida water conservation area. *European Journal of Soil Science*, **60**, 757–769.
- 547 Marchant, B.P., Saby, N.P.A., Lark, R.M., Bellamy, P.H., Jolivet, C.C. & Arrouays,  
548 D. 2010. Robust prediction of soil properties at the national scale: Cadmium  
549 content of French soils. *European Journal of Soil Science*, **61**, 144–152.
- 550 Matérn, B. 1960. Spatial variation. Meddelanden från Statens Skogsforskningsinsti-  
551 tut, 49, No. 5. [2nd Edition (1986), Lecture Notes in Statistics, No. 36, Springer,  
552 New York].
- 553 Newell, E. & Watts, S. 1996. The environmental impact of industrialization in South  
554 Wales in the Nineteenth century: ‘Copper smoke’ and the Llanelli Copper Com-  
555 pany. *Environment and History*, **2**, 309–336.
- 556 Papritz, A. 2007. Robust universal kriging. *Pedometrics 2007*, Tuebingen, Germany  
557 p. 15.
- 558 Rawlins, B.G., Lark, R.M., O’Donnell, K.E., Tye, A.M. & Lister, T.R. 2005. The  
559 assessment of point and diffuse metal pollution from an urban geochemical survey  
560 of Sheffield, England. *Soil Use and Management*, **21**, 353–362.
- 561 Reimann, C., Filzmoser, P. & Garrett, R.G. 2005. Background and threshold: critical  
562 comparison of methods of determination. *Science of the Total Environment*, **346**,  
563 1–16.

- 564 Smith, E., Naidu, R., Weber, J. & Juhasz, A. L. 2008. The impact of sequestration  
565 on the bioaccessibility of arsenic in long-term contaminated soils. *Chemosphere*,  
566 **71**, 773–780.
- 567 Soil Survey of England and Wales 1983. *Soils of Wales*. Ordnance Survey for the Soil  
568 Survey of England & Wales, Southampton.
- 569 Webster, R. & Oliver, M.A. 2007. *Geostatistics for Environmental Scientists*. 2nd  
570 Edition. John Wiley & Sons, Chichester.
- 571 Zhao, Y.C., Xu, X.H., Huang, B., Sun, W.X., Shao, X.X., Shi, X.Z. & Ruan, X.L.  
572 2007. Using robust kriging and sequential Gaussian simulation to delineate the  
573 copper- and lead-contaminated areas of a rapidly industrialized city in Yangtze  
574 River Delta, China. *Environmental Geology*, **52**, 1423–1433.
- 575 Zhang, C., Luo, L., Xu, W. & Ledwith, V. 2008. Use of local Moran's I and GIS to  
576 identify pollution hotspots of Pb in urban soils of Galway, Ireland. *Science of*  
577 *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<sup>-1</sup> unless stated.

Element	As		Cu		Fe <sub>2</sub> O <sub>3</sub> (%)		Pb		Sn	
	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	
<sup>a</sup> P-value	0.67		0.65		0		0.78		0.41	

<sup>a</sup> P-value for null hypothesis that variable is independent of Fe<sub>2</sub>O<sub>3</sub>.

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

	Cu	As	Pb	Sn
${}^a\bar{\theta}_M$	1.15	1.03	0.88	0.97
$\check{\theta}_M$	0.35	0.39	0.30	0.40
Estimator	Dowd	Genton	Dowd	Dowd
${}^b\bar{\theta}_R$	1.40	1.19	1.03	1.15
$\check{\theta}_R$	0.44	0.46	0.41	0.44
$c$	2.1	2.3	2.7	2.4
${}^c\bar{\theta}_c$	1.01	1.01	1.00	1.00
$\check{\theta}_c$	0.40	0.44	0.41	0.44

${}^a\theta_M$  cross-validation statistic for Matheron estimator

${}^b\theta_R$  cross-validation statistic for best robust estimator

${}^c\theta_c$  cross-validation statistic for winsorized data

**Table 3** Land use (current and historic) types for point-source metal and metalloïd contaminants (soil concentration in mg kg<sup>-1</sup>). References correspond to labels in Figure 6. Features next to land use (derived from Ordnance Survey maps) shown in parentheses.

Ref.	Land use at given date		
	Concentration	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:

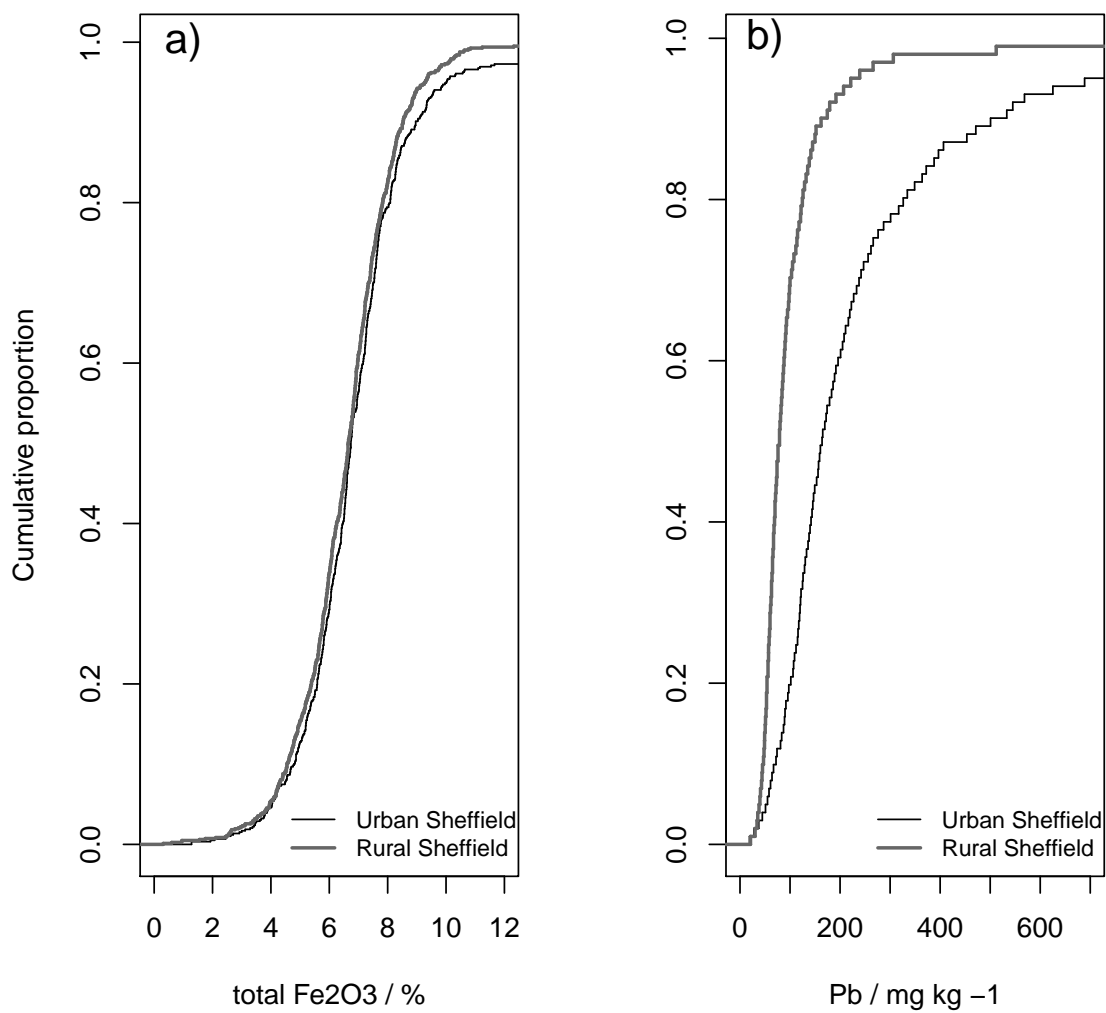


Figure 2:

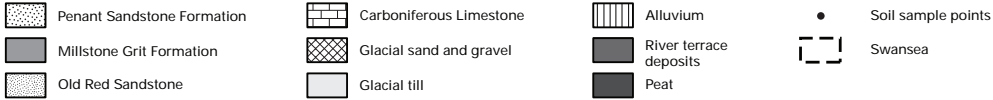
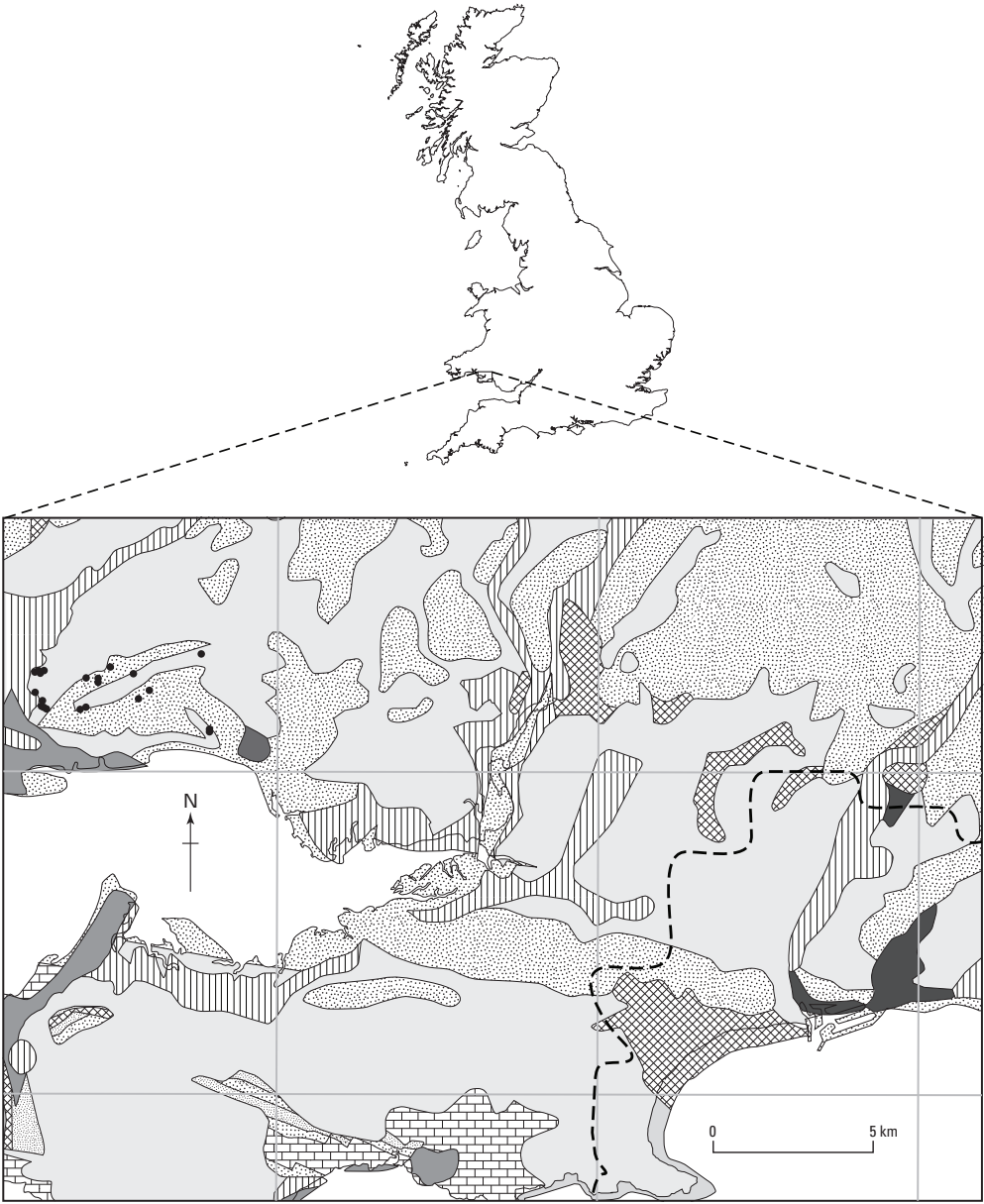


Figure 3:

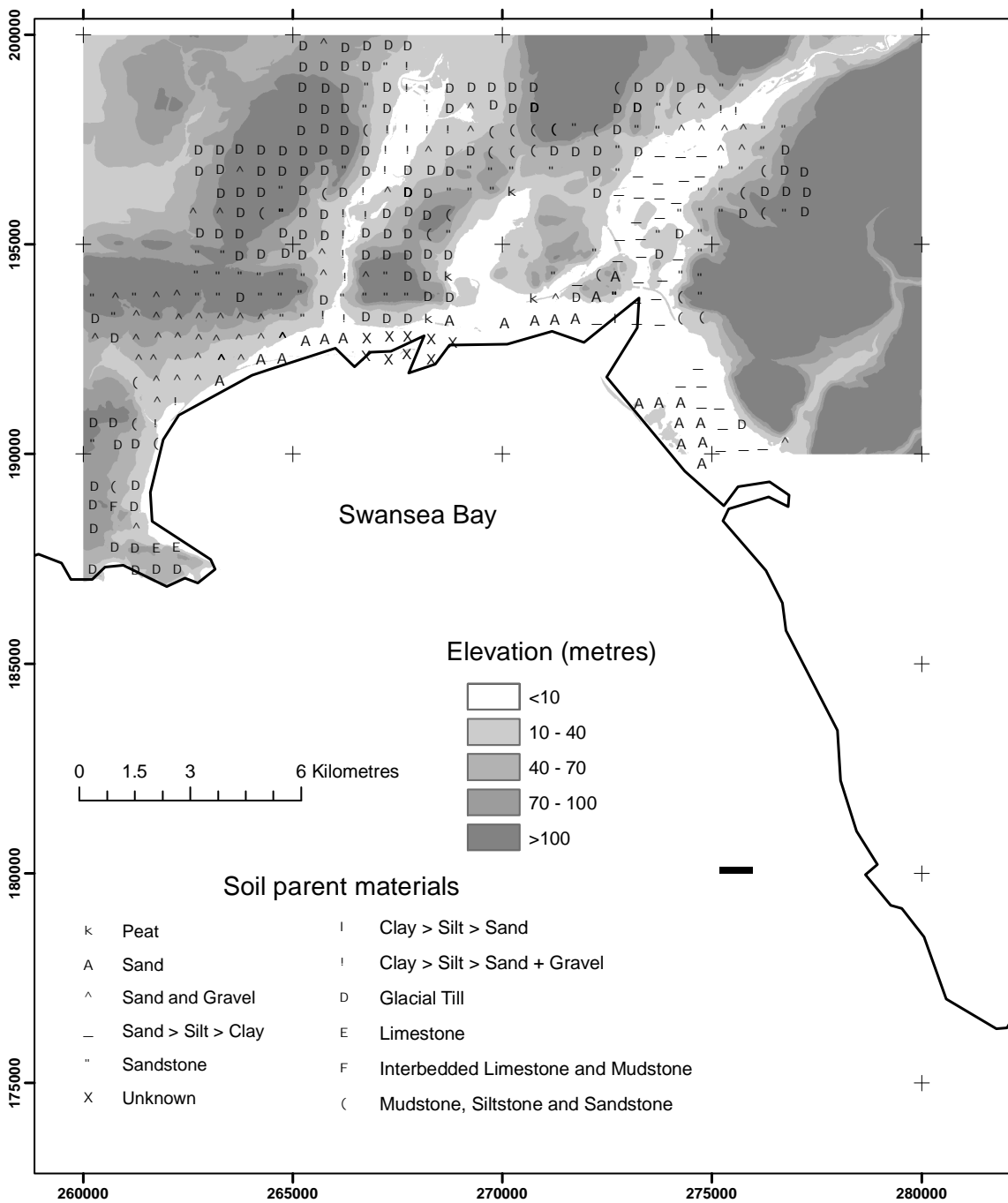




Figure 4:

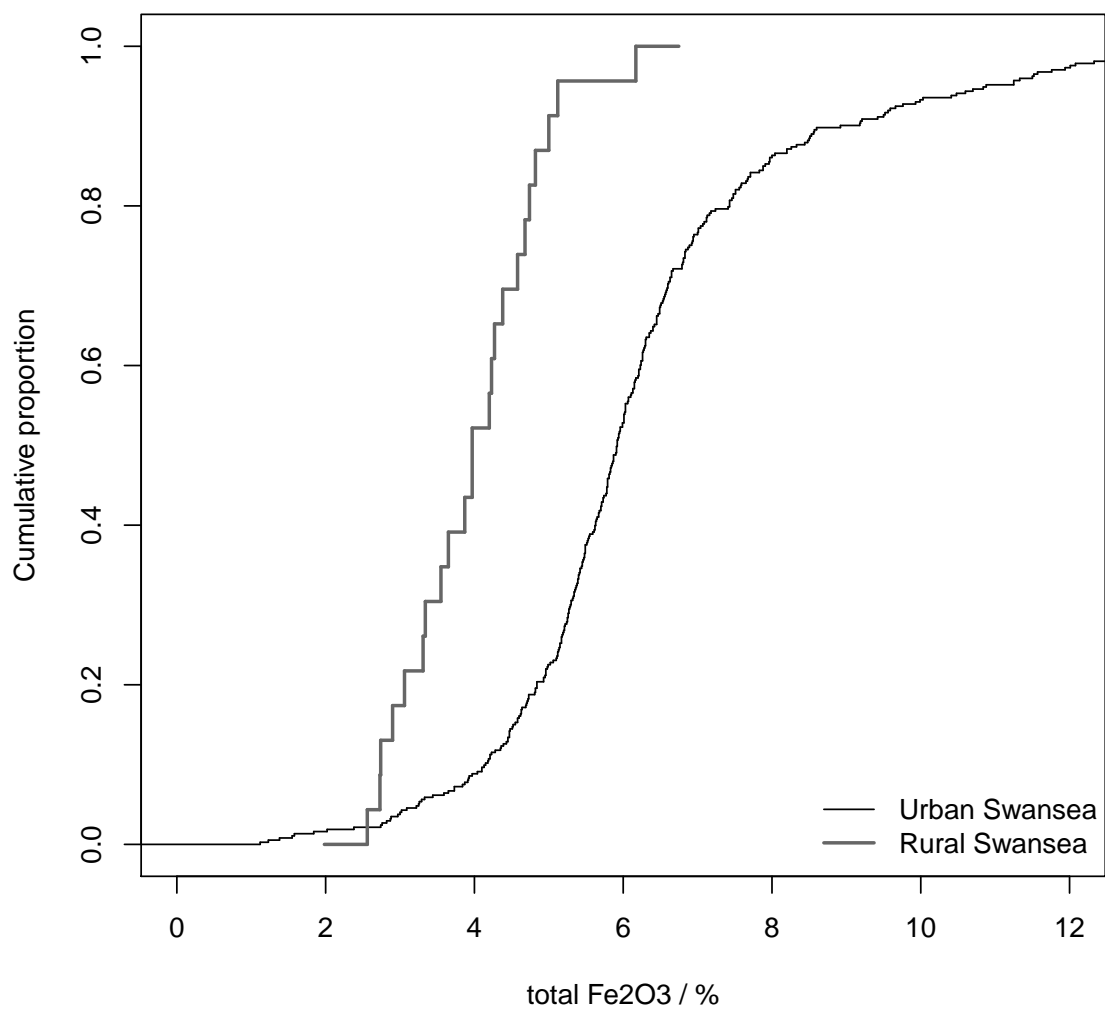


Figure 5:

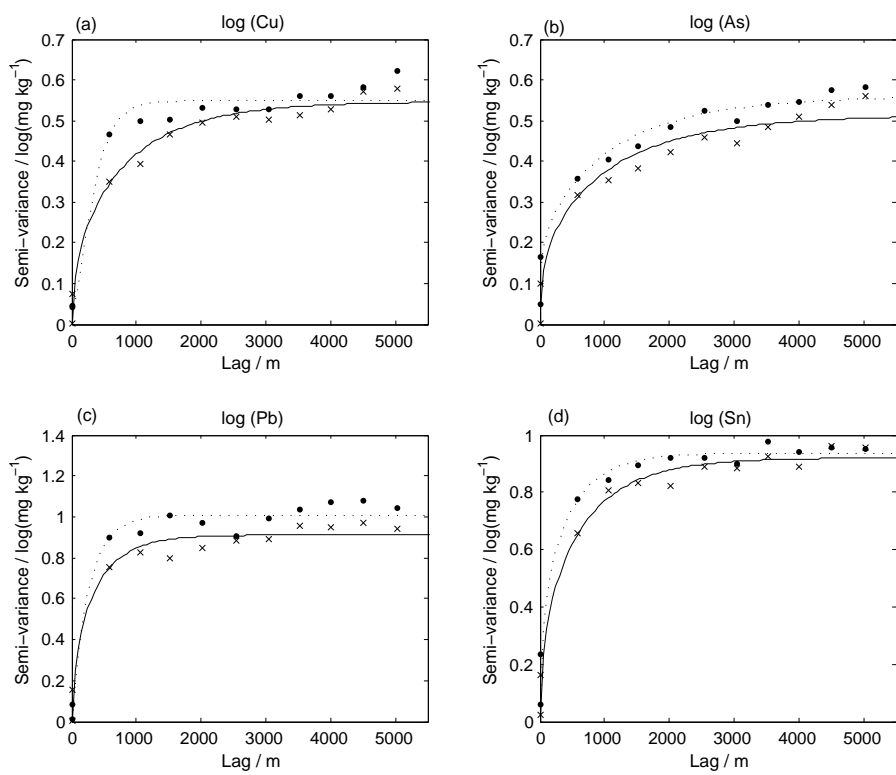


Figure 6:

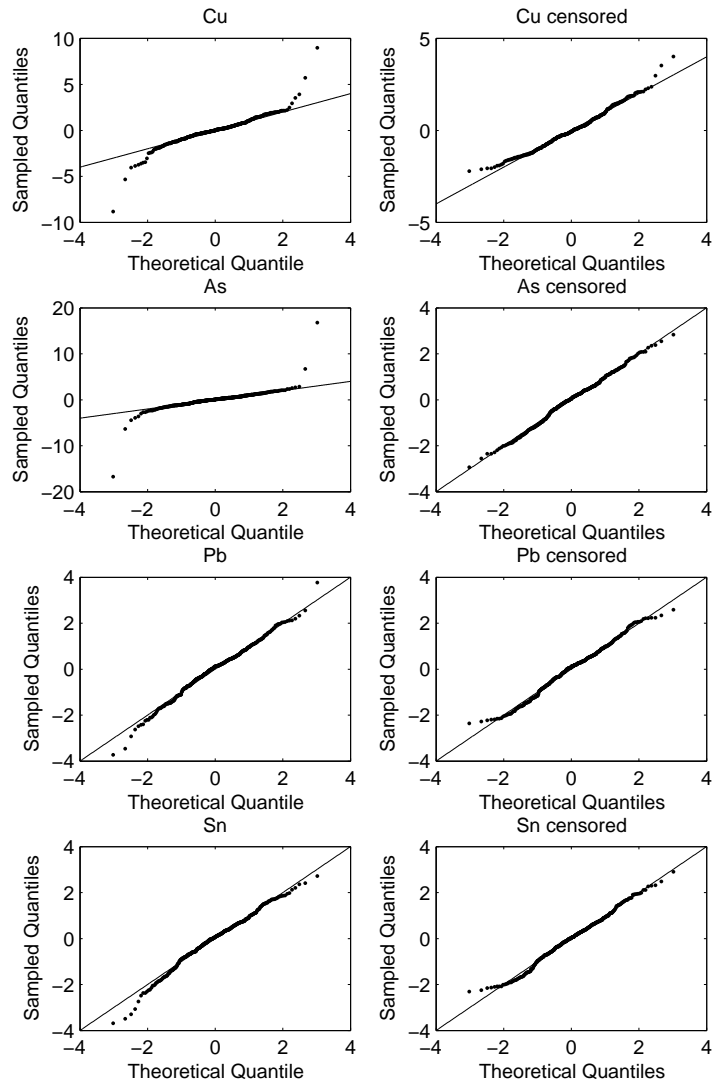


Figure 7:

