The effect of Lead in soil on Crime Deprivation in Derby, Leicester and Nottingham

Mark Cave[a*], Joanna Wragg[a], Robert Lister[a]

[a]British Geological Survey, Keyworth, Nottingham, NG12 3NX

mrca@bgs.ac.uk

*Corresponding Author, British Geological Survey, Keyworth, Nottingham, NG12 3NX, e-mail: mrca@bgs.ac.uk

Abstract

The aim of this study was to test the hypothesis that soil Pb is associated with criminality in selected urban environments within the UK. The study used geological and geochemical information and soil Pb data from Derby, Leicester and Nottingham, collected as part of a geochemical survey of urban soils. Crime and other associated socio-economic data were provided by a national survey on deprivation in the UK. The data were modelled using crime deprivation as the dependant variable and Pb, Sn and Ce in soil as well as three socio-economic factors associated with personal deprivation, population density and environmental deprivation as predictor variables. Both the generalised linear and the random forest modelling strategies showed that the socio-economic predictor variables and spatial associations were important in predicting crime deprivation. Pb and the two other soil chemistry parameters (Sn and Ce) were not important predictors of crime deprivation in Leicester and Nottingham. Pb and its interactions with spatial and socio-economic factors were, however, shown to have a significant effect on crime deprivation in Derby. The random forest model for Derby showed that there was an antagonistic interaction effect between Pb in soil and personal deprivation. The random forest model was used to produce “dose-response”

curves of the effect of Pb in soil on crime deprivation under different spatial and socio-economic conditions.

Keywords

Pb; crime; soil; urban; random forest

1. Introduction

Lead (Pb) is toxic to humans with one of its main detrimental effects being neurological, behavioural and development impairment in children (ATSDR 2007). There is an increasing body of evidence that suggests that if children are exposed to Pb at a young age then the impairment of their neurological development can lead to a propensity for them to commit crime as young adults (i.e. about 20 years later) (Bellinger, 2008; Marcus et al., 2010; Mielke and Zahran, 2012; Nevin, 2007; Taylor et al., 2016; Wright et al., 2008). Some of these studies suggest that that the exposure comes from lead in air particulates (Mielke and Zahran, 2012; Taylor et al., 2016). Other studies (Laidlaw and Taylor, 2011; Young et al., 1992; Zahran et al., 2010) show that exposure to Pb, as measured by blood Pb, is also controlled by the Pb content of local soil. The aim of this study is to test the hypothesis that soil Pb is associated with criminality in selected urban environments within the UK. The study used soil Pb data collected as part of a geochemical survey of urban soils and crime and other associated socio-economic data provided by a national survey on deprivation in the UK.

2. Methodology

Three cities in the Midlands of the UK have been chosen as test locations, these are Derby, Leicester and Nottingham (Figure 1). Whilst all three cities are relatively close they have different industrial histories and natural geochemical background

signatures. The details of the industrial histories, geochemistry of soils and the geology of these three cities has been previously published (Scheib and Nice, 2008) but to summarise these in brief:

Figure 1 Location of Derby, Leicester and Nottingham within the UK

Derby

Derby city in the East Midlands lies on the banks of the River Derwent and southern Derbyshire. In the census of 2001 the population was recorded at just under 222,000. Derby and Derbyshire played a pivotal role in Britain's industrial revolution. In 1771, Derby was the site of the first water powered silk mill in Britain, positioned on the banks of the River Derwent. Beginning of the 19th century saw Derby emerging as an engineering centre manufacturing machine tools. In 1840, Midland Railway set up its works and headquarters in Derby. It continued to be a significant railway centre, hosting both British Rail workshops and research facilities. Although

much less important than in years gone by, train manufacture continues in Derby and Derby Midland Station retains an important strategic role in the rail network. Derby's two biggest employers are Rolls-Royce plc (which has been established in the city since 1907) and the Toyota Motor Corporation (in operation since 1992), although this lies to the southwest of the sampling area. Rolls Royce originally manufactured cars until 1946, when motor production was transferred to Crewe. Today, Rolls Royce Derby concentrates on aeronautical and marine engineering.

Figure 2 shows the superficial geology of Derby and the soil sampling sites. Apart from a small region in the north, the entire sampling area is underlain by the Triassic Mercia Mudstone Group. Its deposition reflects a complex mixture of mainly continental environments in which thick sequences of red-brown or rarely green-grey mudstone of aeolian and lacustrine origin accumulated, punctuated by fluvial episodes that deposited beds of grey-green dolomitic silt and sandstone, commonly referred to as "skerries".

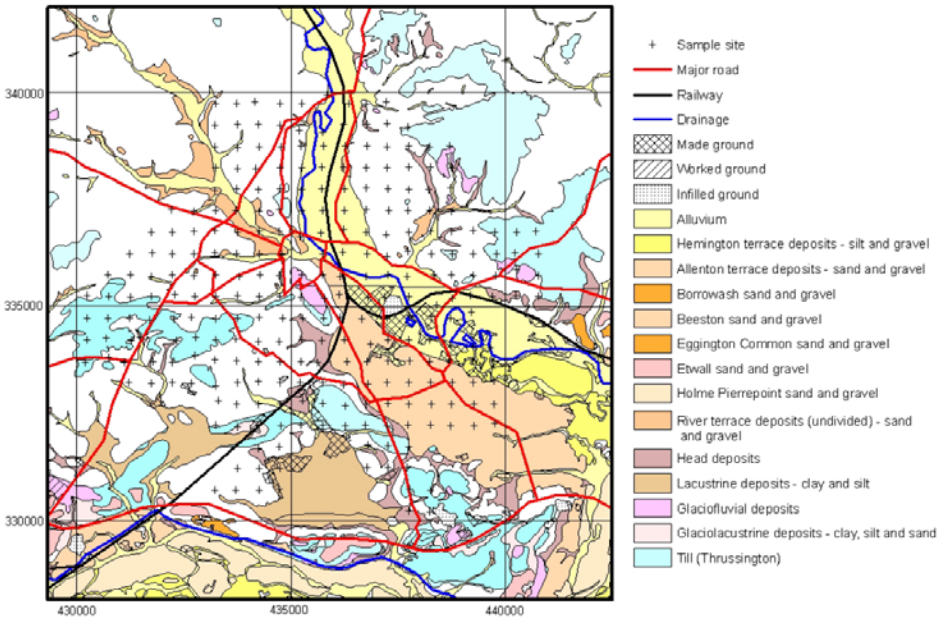

Figure 2 Superficial geology map of Derby (1:50 000 British Geological Survey©)

A major Quaternary feature of the urban area of Derby is the floodplain alluvium of the River Derwent and its tributaries.

Leicester

Leicester is the most populated city in the East Midlands. It lies in the River Soar valley at the eastern edge of the National Forest. In 2004 the population of the city was approximately 285 000 with a further 330 500 living in the surrounding urban areas. A rapid industrialisation began with the construction of the Grand Union Canal in the 1790s, linking Leicester to London and Birmingham. From then Leicester developed from a traditional market town into an industrial centre during the nineteenth century. By 1832 railways had arrived and, with the opening of the Leicester and Swannington line, they provided a supply of coal to the town from nearby collieries. This led to the establishment of industrial complexes particularly along the riverside, which included engineering works and factories manufacturing boots, hosiery and knitwear. Following the First World War, industrial estates were established, to separate industry from the residential areas that were also being developed and expanded to house the growing population. Major industries in Leicester today include food processing, hosiery, footwear, knitwear, engineering, electronics, printing and plastics.

Figure 3 shows the superficial geology and soil sampling locations showing that the sampling area of Leicester is almost completely covered by superficial deposits. Alluvial and river terrace deposits dominate the areas along the Rivers Sence and Soar and till underlies much of the urban area, with sporadic small outcrops of glaciofluvial deposits. Made ground is likely to be extensive across the urban area but is not well mapped.

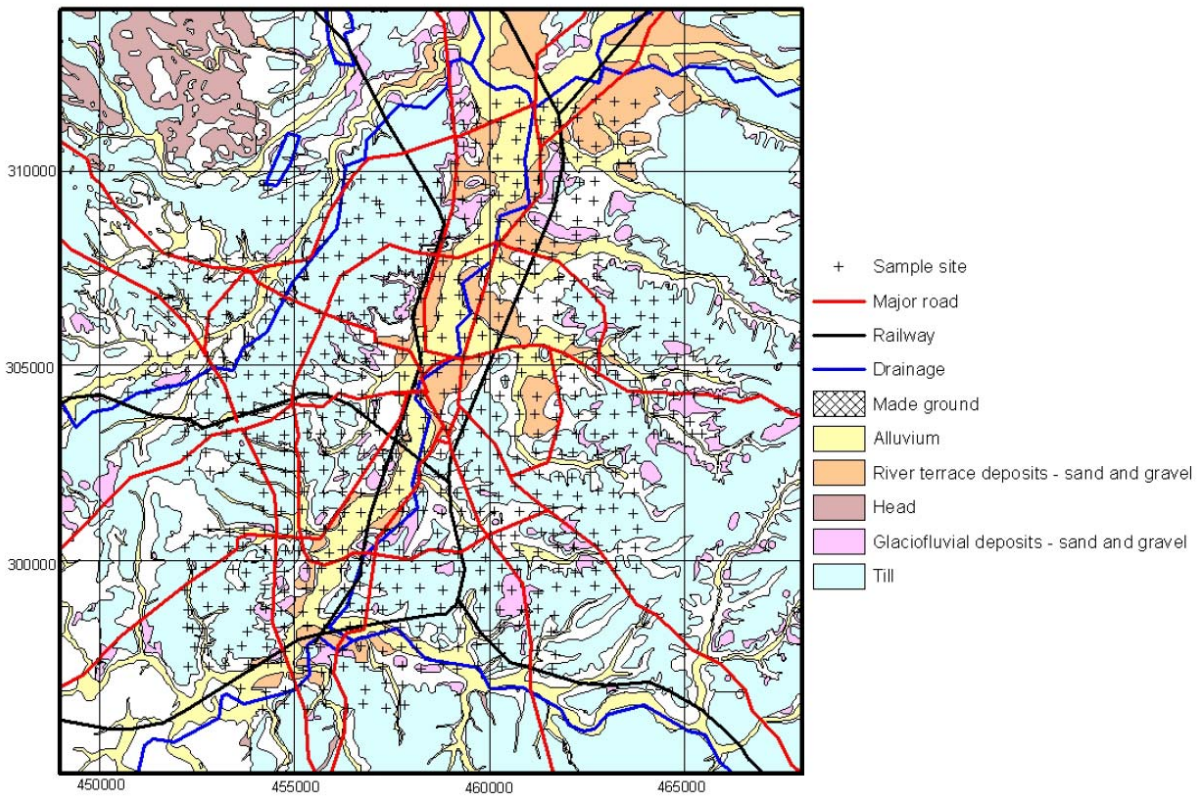

Figure 3 Superficial geology map of Leicester (1:50 000 British Geological Survey©)

Nottingham

Nottingham is the county town of Nottinghamshire in the East Midlands. It lies on the River Leen, but is more commonly associated with the River Trent, of which the River Leen is a tributary. The 2001 census recorded a population of 270 300 in Nottingham itself, with an additional 613 000 living in the surrounding area of Greater Nottingham. Nottingham is most famous for its lace-making industry and other famous industries within the city are Boots, which was founded in 1849 and is now based in the Beeston area. The tobacco company John Player & Sons was based in Nottingham for nearly 150 years but closed in 2016. Until recently, bicycle manufacturing was a major industry, with Nottingham being the birthplace of Raleigh Cycles in 1886. The factory has since been demolished to make way for the expansion of the University of

Nottingham campus. Of the three cities Nottingham's industrial history has been associated with lighter industries compared to Derby and Leicester.

Nottingham owes its location to its geology, where the topographically resistant Triassic sandstone has survived as a bluff of high and dry ground that overlooks a shallow crossing point of the River Trent. The solid geology of Nottingham features a diverse stratigraphical succession consisting of Carboniferous, Upper Permian, Triassic and Jurassic rocks. The succeeding Mercia Mudstone Group of predominantly red mudstone with sporadic beds of siltstone and sandstone underlies most of Nottingham. The northeast of Nottingham is underlain predominantly by mudstone which is underlain in turn by sandstone and siltstone. The southern half of Nottingham is dominated by a sequence of mudstones, sandstones and siltstones.

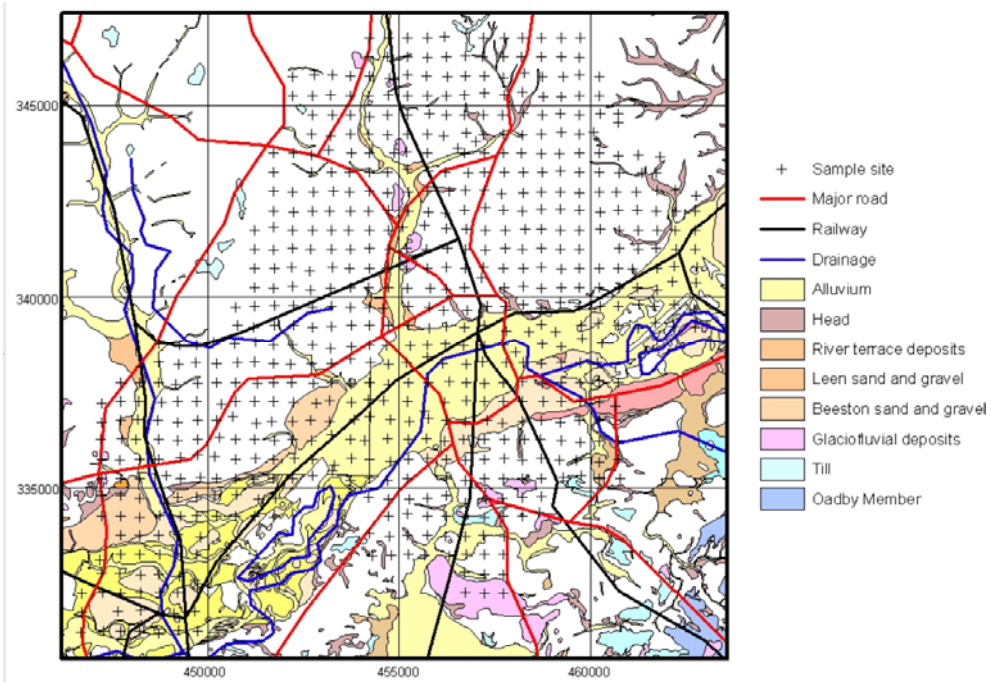

Figure 4 Superficial geology map of Nottingham (1:50 000 British Geological Survey©)

The major superficial deposits in the Nottingham sampling area (Figure 4) occur along the valleys of the River Trent and Erewash and their tributaries, and comprise alluvial sand and gravel deposits.

2.1 Soil data

The soils from the three cities were collected as part of the BGS Geochemical Baseline Survey of the urban Environment (G-BASE) (267 samples for Derby, 275 for Leicester and 284 for Nottingham). Topsoil samples were collected at a depth of ca. 5-20 cm from open ground on a 500 m grid at a density of approximately 4 samples per km^2. At each site, composite samples, based on 5 sub-samples taken at the centre and four corners of a 5 m square were collected from the topsoil (5-20 cm depth). Forty-eight chemical elements were determined in the <2 mm size fraction of the topsoils using X-ray fluorescence spectrometry (XRFS), together with loss on ignition (LOI at 450 °C) and pH. Sample preparation, analytical methods, and quality control procedures have been previously described (Allen et al., 2011; Johnson, 2011).

2.2 Crime and socio economic related data

The English Indices of Deprivation (EID) (Smith et al., 2015) is a freely available data set containing a combination of socio-economic data covering all of England summarised into areas known as Lower Super Output Areas (LSOA). These are geographic areas for the collection and publication of small area statistics. They have an average of roughly 1,500 residents and 650 households. Measures of proximity (to give a reasonably compact shape) and social homogeneity (to encourage areas of similar social background) are also included. There are 151 LSOAs covering Derby, 192 covering Leicester and 182 covering Nottingham. For each LSOA the EID provides the following data on deprivation indices related to Income, Employment,

Health, Education, Crime, Housing and Living environment. A detailed description of the way in which these deprivation indices are constructed has been previously described (Smith et al., 2015) but in brief:

The Income Deprivation Domain measures the proportion of the population in an area experiencing deprivation relating to low income. The definition of low income used includes both those people that are out-of-work, and those that are in work but who have low earnings (and who satisfy the respective means tests).

The Employment Deprivation Domain measures the proportion of the working-age population in an area involuntarily excluded from the labour market. This includes people who would like to work but are unable to do so because of unemployment, sickness or disability, or caring responsibilities.

The Health Deprivation and Disability Domain measures the risk of premature death and the impairment of quality of life through poor physical or mental health. The domain measures morbidity, disability and premature mortality but not aspects of behaviour or environment that may be predictive of future health deprivation.

The Education, Skills and Training Domain measures the lack of attainment and skills in the local population. The indicators fall into two sub-domains: one relating to children and young people and one relating to adult skills.

The Crime Domain measures the risk of personal and material victimisation at local level. It uses: the rate of violence per 1,000 at-risk population; the rate of burglary per 1,000 at-risk properties; the rate of theft per 1,000 at-risk population; and the rate of criminal damage per 1,000 at-risk population.

The Barriers to Housing and Services Domain measures the physical and financial accessibility of housing and local services. The indicators fall into two subdomains: ‘geographical barriers’, which relate to the physical proximity of local services, and ‘wider barriers’ which includes issues relating to access to housing such as affordability.

The Living Environment Deprivation Domain measures the quality of the local environment. The indicators fall into two sub-domains. The ‘indoors’ living environment measures the quality of housing; while the ‘outdoors’ living environment contains measures of air quality and road traffic accidents.

2.3 Comparison of cities

Figure 5 show a comparison between the Pb in soil samples from the three cities as a box and whisker plot and as an overlaid probability density plot.

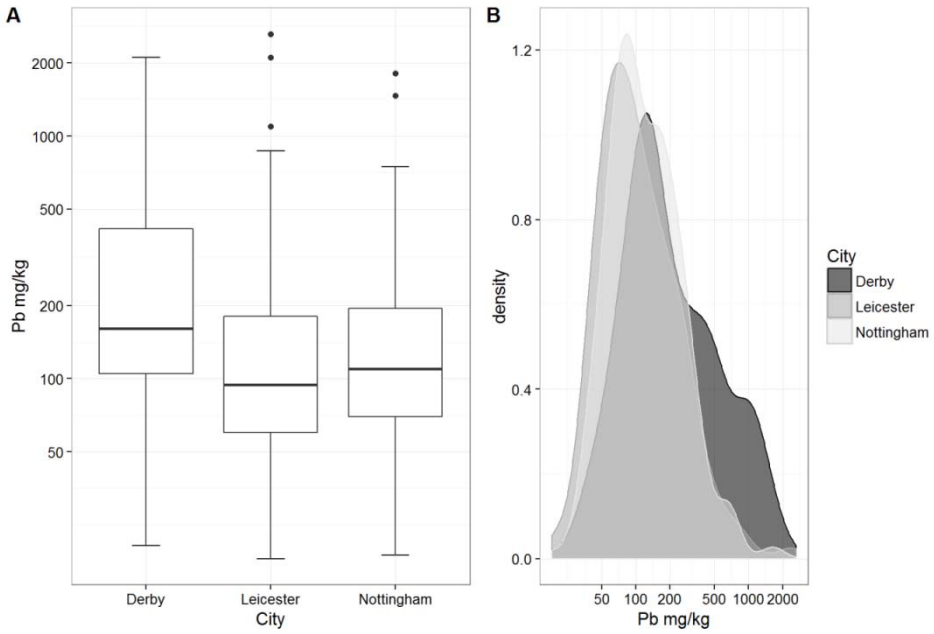

Figure 5 Comparisons of Pb concentration in soil between the three cities. A-Boxplot, B-probability density plot

Derby clearly has a higher median Pb values with an increased proportion of samples above 500 mg/kg. Figure 6 shows the spatial distribution of Pb concentrations in the soil samples for the three cities.

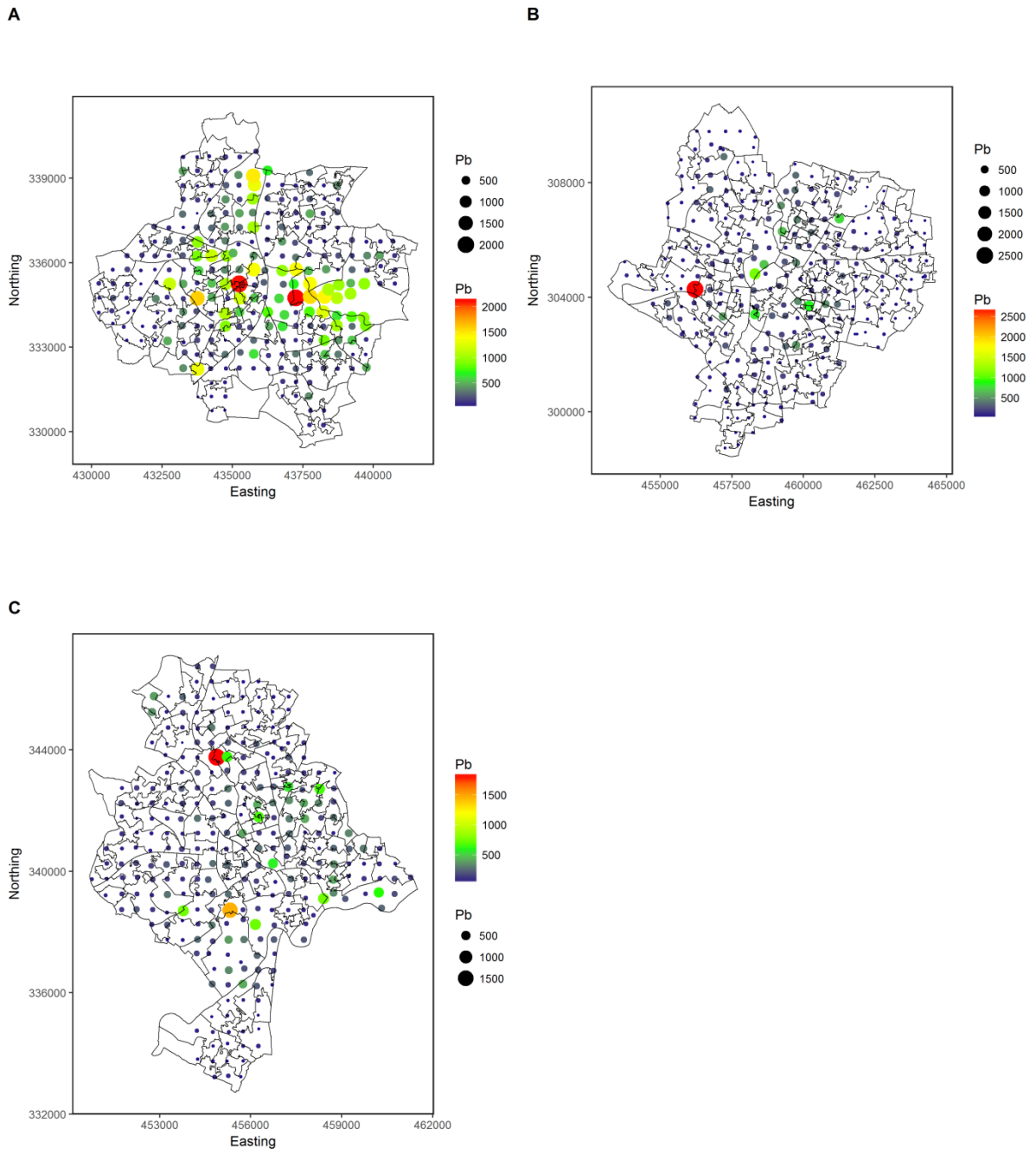

Figure 6 Pb concentrations at each sampling point for each city. A Derby; B Leicester; C Nottingham

Comparing the Pb data distributions (Figure 5) and the spatial distributions (Figure 6) with the superficial geology maps of the three cities (Figure 2, Figure 3, Figure 4)

provides some insight into the controlling factors Pb in the soils. For both Nottingham and Leicester the highest concentrations of Pb are found towards centre of the cities (more on the east side for Nottingham) and do not show any clear spatial relationship to the underlying geology, suggesting that the Pb is derived from diffuse anthropogenic sources. For Derby, however, there is a clear pattern in the Pb distribution forming a broad arc starting from the north of the city down to the city centre and then carrying on through to the east of the town (Figure 6 A). This coincides with the alluvium and sand and gravel deposits in the flood plain of the river Derwent. The River Derwent flows through the Peak District to the North of Derby where there are high concentrations of Pb in the rocks soils arising from Pb mineralisation in the carboniferous limestones (Ander et al., 2013; Rawlins et al., 2012). This suggests one of the main sources of Pb in the Soils of Derby come from river transported of Pb rich material from the Peak District.

The soil data were joined to the socio-economic data by attributing soils within a given LSOA to the EID data associated with that LSOA. Figure 7 compares the crime deprivation for the three cities showing that Leicester and Derby are broadly similar with Derby having a lower median value. Figure 8 shows box and whisker plots of Pb in the three cities split into increasing crime deprivation categories (reading left to right and top to bottom).

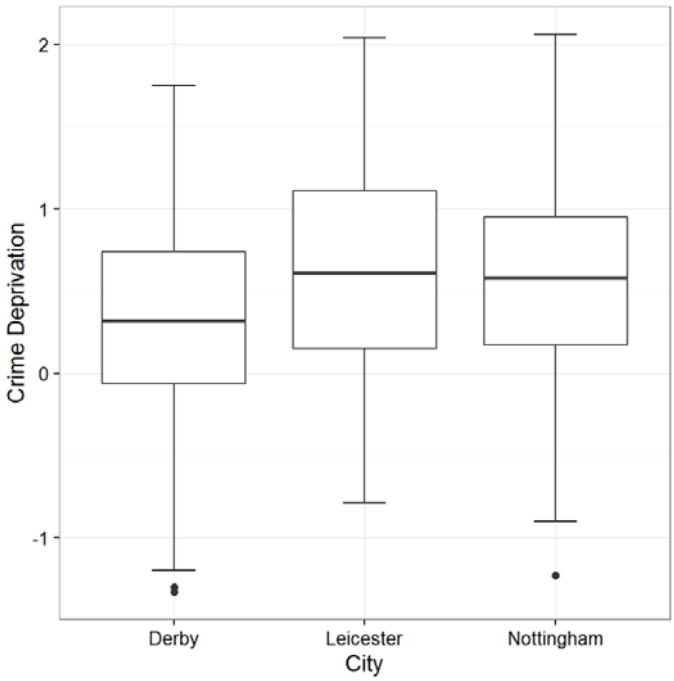

Figure 7 Comparison of crime deprivation between cities

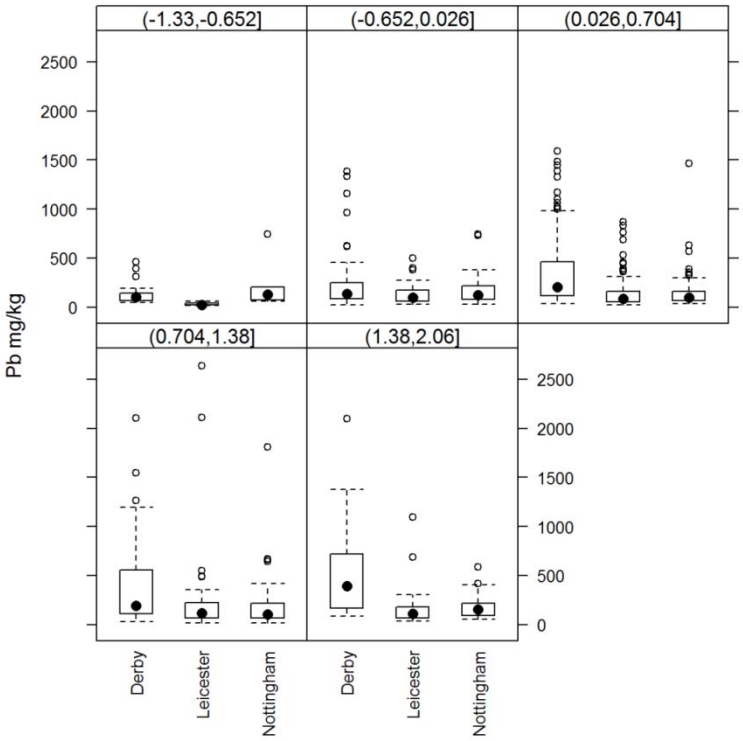

Figure 8 Box and whisker plot of Pb in soil partitioned into crime deprivation categories by city

For the two lowest crime categories the Pb boxplots are similar for all three cities. As the crime deprivation increases in the last three categories, however, the Leicester and Nottingham Pb distributions remain similar and low but Derby shows increased Pb. This suggests that there is a link between higher Pb concentration in soil and higher crime deprivation in Derby.

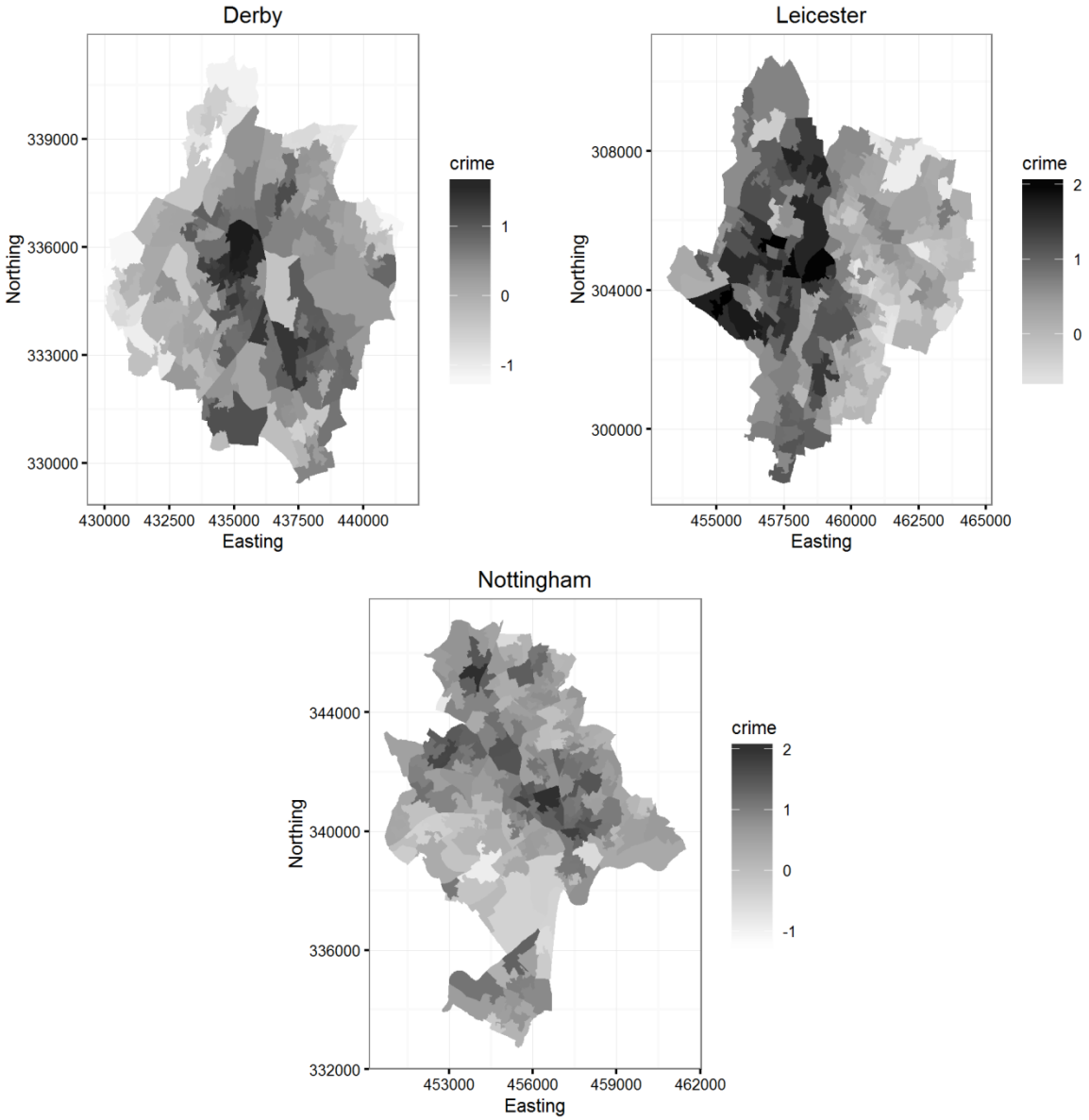

Figure 9 Crime deprivation summarised by LSOA in the 3 cities

Figure 9 shows the spatial distribution of crime deprivation in the three cities. For Derby and Nottingham the highest crime deprivation tends to be in the centre of the city whereas Leicester has a clear east west split with higher crime deprivation on the west side. There does not appear to be any clear visual association between the Pb spatial distribution (Figure 6) and the crime deprivation.

2.4 Data Pre-processing

As the model involves combining geochemical, socio-economic and demographic data sources, all data were transformed to approximate normal distributions using the Yeo-Johnson algorithm (Yeo and Johnson, 2000) followed by mean centering and scaling the data. This was preferred over the BoxCox method since some of the deprivation indices have negative values.

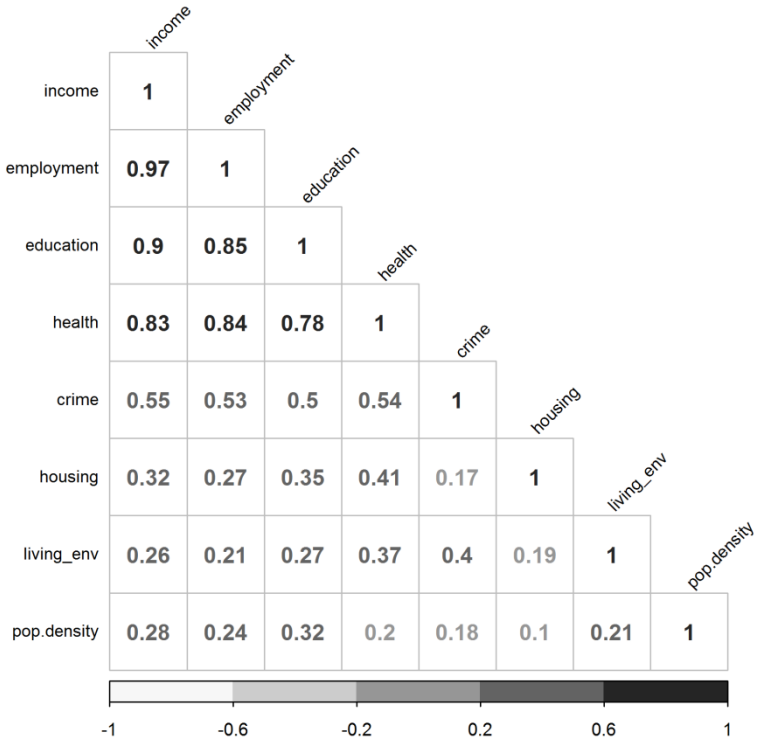

Figure 10 Correlation between socio-economic factors in Derby, Leicester and Nottingham

All data transformations were carried out using the “caret” package in the R programming language (Kuhn, 2016).

The EID different domains of deprivation provide a broad spectrum picture of socio-economic descriptors which could also have an effect on the crime deprivation and therefore act as suitable confounder covariates for soil Pb. An examination of the correlations between these covariates (Figure 10) shows that that income, employment, education and health are highly correlated (Pearson correlations> 0.78).

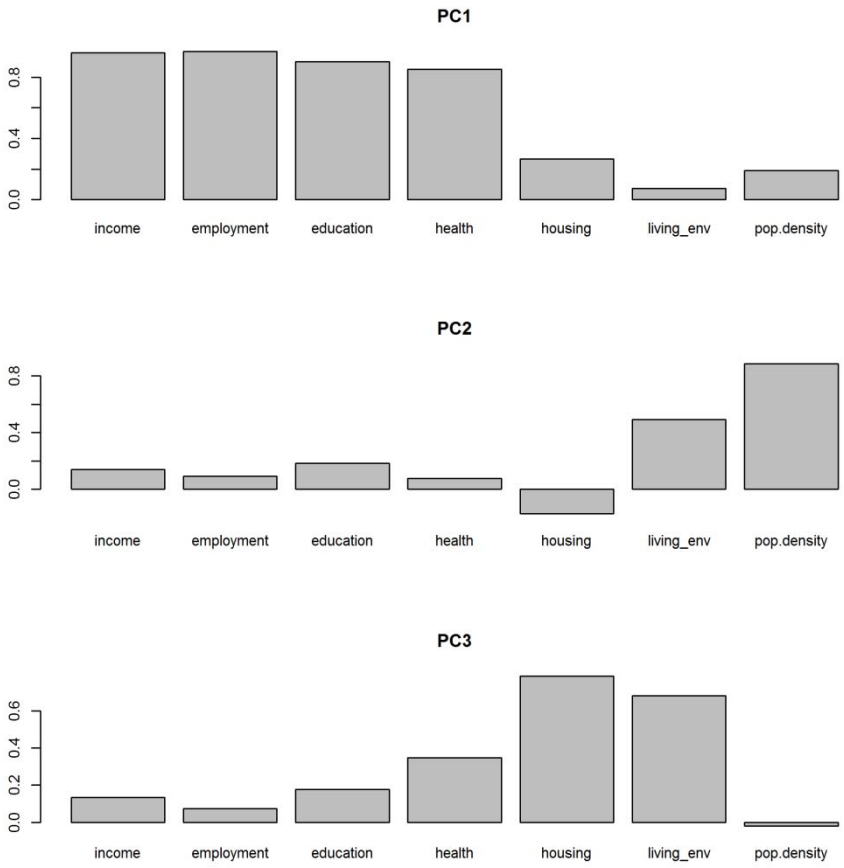

Figure 11 Deprivation loadings on the 3 principal components used to model the 2015 Deprivation indices for Derby, Leicester and Nottingham.

For modelling purposes and to help in the final model interpretation it is better not to have highly correlated predictor variables which can cause unstable model outcomes

and make interpretation of the underlying causation less clear. To achieve this, the deprivation indices data was subjected to principal component analysis followed by varimax rotation, allowing the 6 deprivation indices and population density to be reduced to three orthogonal principal components (PCs). The loadings of the 7 variables are shown in Figure 11. PC1 has high loadings on Income, Employment, Health, Education and Health; all factors associated with personal circumstances, and has been named “Personal Deprivation” (PD). PC2 has its highest loading on Population Density and to a lesser extent Living Environment and therefore takes on the name of “Population Density” (Pop.Dens). PC3 has its highest loading on Housing and to a lesser extent Living Environment and Health which are more related the environment and takes on the name of “Environment deprivation” (ED).

2.5 Data modelling

The aim was to see if Pb in soil could be shown to be the cause of some aspects of crime deprivation in the three cities. The modelling, however, needs to take into account other factors that are likely to be important causal factors associated with crime (Wikström et al., 2012) to ensure that the effect of Pb alone can be isolated and that Pb in soil is not acting as a proxy for other underlying variables. The modelling therefore used the CD as the dependant variable and Pb as a predictor variable alongside 5 other potential confounder variables. These consisted of the three socio-economic derived factors from the EID data, PD, Pop.dens and ED alongside two other soil chemistry variables. The two additional soil chemistry parameters were; Sn which is known to be a an indicator of anthropogenic inputs in urban environments because its natural concentrations are usually low (<5mg/kg) and Sn has a low geochemical mobility (common sources of Sn are old paint, glazed pottery, electrical solder and tinplate/old tin cans)(Albanese and Breward, 2011); and Ce which provides

a measure of natural background (Aide and Aide, 2012) it has been shown that Ce in urban environments in the UK has no systematic variation from estimated upper crustal concentrations (Flight and Scheib, 2011).

Maps of the spatial distribution Ce and Sn in the individual soil samples in the three cities are shown in the SI (Figures S1 and S2).

Keeping in mind the quotation from Box and co-authors "Essentially, all models are wrong, but some are useful." (Box and Draper 1987), two independent modelling approaches were used to provide a more robust overview of the potential relationship between Pb in soil and crime. The two approaches were mixed effect modelling and machine learning.

2.5.1 Generalised least squares modelling

The relationship between crime deprivation and potential predictor variables Ce, Pb , Sn and the three socio-economic derived factors PD, ED and Pop.D were explored using generalised least squares (GLS) modelling using the “nlme” package in the R programming language and its associated “gls” command (Pinheiro et al., 2016). The procedure for model selection follows that given in (Zuur et al., 2009) and takes the form of the following steps:

i) Check to see if there is spatial autocorrelation in the model and if so select the best autocorrelation fit using the Akaike Information criterion (Akaike, 1974) as the selection criterion.
ii) Use the likelihood ratio test to check to see if interaction effects between Pb and the socio-economic factors are significant.

iii) Check the final model for violation of model assumptions (normal distribution and homogeneity of residuals with no spatial autocorrelation; collinearity as measured by the variance inflation factor (VIF) (Vuong, 1989) is <2), the fraction of variance in the crime deprivation accounted for by the model and the relative importance of the predictor variables.

The outputs relating to the model selection process in steps i)-iii) are given in the Supplementary Information (SI).

2.5.1.1 Derby

Table S1 in the SI shows that inclusion of spatial correlation produces lower AIC values and that the exponential spatial correlation fit (Pinheiro et al., 2016) has the lowest AIC. The variogram fit to the raw data is shown in Figure S4 in the SI. The likelihood ratio test as implemented by the "anova" function in the R programming language (R Core Team, 2016) shows that an inclusion of a first order interaction between Pb and the three socio-economic factors as predictor variables gives a significant improvement in the model fit (p=0.0079). The GLS, with the inclusion of Pb interaction effects, shows minimal collinearity of predictors (VIF values for all predictors < 1.6) and that there is no spatial correlation in the standardised residuals (Figure S5 in the SI). In addition, the residuals are homoscedastic with an approximate normal distribution (Figure S6 in the SI). The table of coefficients for the optimised GLS model for Derby (Table S2 in the SI) shows that PD is the only direct effect predictor significant at p<0.05, but the interactions of Pb with PD and with Pop.d are both significant. The model explains 53% (p-value 0.05) of the variance in the crime deprivation data for Derby.

2.5.1.2 Leicester

Table S3 in the SI shows that inclusion of spatial correlation produces lower AIC values and that the spherical spatial correlation fit (Pinheiro et al., 2016) has the lowest AIC. The variogram fit to the raw data is shown in Figure S7 in the SI. The likelihood ratio test shows that an inclusion of a first order interaction between Pb and the three socio-economic factors as predictor variables is not a significant improvement in the model fit (p= 0.302). The GLS, without the inclusion of Pb interaction effects, shows collinearity between Pb and Sn (Pearson correlation coefficient of 0.83). Removal of Sn from the model gives VIF values for the remaining predictors all <1.15. This model gives no spatial correlation in the standardised residuals (Figure S8 in the SI). In addition, the residuals are homoscedastic with an approximate normal distribution (Figure S9 in the SI). The table of coefficients for the optimised GLS model for Leicester (Table S4 in the SI) shows that the three socio-economic variables are all significant but the soil parameters (Ce and Pb) are not. The model explains 12% of the variance in the crime deprivation data for Derby.

2.5.1.3 Nottingham

Table S5 in the SI shows that inclusion of spatial correlation produces lower AIC values and that the rational quadratics spatial correlation fit (Pinheiro et al., 2016) has the lowest AIC. The variogram fit to the raw data is shown in Figure S10 in the SI. The likelihood ratio test shows that an inclusion of a first order interaction between Pb and the three socio-economic factors as predictor variables is not a significant improvement in the model fit (p=0.114). The GLS, without the inclusion of Pb interaction effects, shows collinearity between Pb and Sn (Pearson correlation coefficient of 0.73). Removal of Sn from the model gives VIF values for the remaining predictors all < 1.16. This model gives no spatial correlation in the standardised residuals (Figure S11 in the SI). In addition, the residuals are homoscedastic with an

approximate normal distribution (Figure S12 in the SI). The table of coefficients for the optimised GLS model for Nottingham (Table S6 in the SI) shows that the three socio-economic variables are all significant but the soil parameters (Ce and Pb) are not. The model explains 38% of the variance in the crime deprivation data for Nottingham.

2.5.1.4 GLS modelling outcomes

The GLS modelling of crime deprivation in the three cities has shown that, unsurprisingly, there is spatial dependence between the sampling locations and that the socio-economic predictor variables are important predictor variables. The soil variables (Ce, Pb and Sn in soil) are, however, not significant predictors of crime deprivation in Leicester and Nottingham but that the interaction of Pb with socio-economic predictor variables in Derby seems to have a significant effect on the crime deprivation. Whilst the optimised GLS models do not show any clear patterns in their residual plots (Figures S6, S9 and S12 in the SI) which suggest that there is non-linearity in the models that is not being accounted for, the variance explained by each model is ca.50 % or less. This may be because an important confounder covariate is not being included or that there are interactions between covariates which have not yet been explored. Finding out the best combination of interaction effects and the best statistical model to apply can be a time consuming and difficult even with the relatively few predictor variables in this model. One approach which automatically takes into account interactions and non-linearity is the machine learning methodology “random forest” (Breiman, 2001b).

2.5.2 Random Forest Modelling

Random forests (RF) are an ensemble learning method for classification, regression and other tasks, which operate by constructing a multitude of decision trees at training

time and outputting the mean prediction (regression) of the individual trees. To reduce the variance and bias between the decision trees the idea of “bagging” was introduced (Breiman, 2001a) in which many large trees are fitted to bootstrap-resampled versions of the training data and, additionally, at each decision node of each tree a random sample the original prediction parameters are used to make the next decision. The “randomForest” package within the R programming language (Liaw and Wiene, 2002) was used to set up a models with Ce, Pb and Sn in soil, the three socio-economic variables and the easting and northing as predictor variables and crime deprivation as the dependant variable for each of the three cities.

The first stage in producing the RF models is to optimise the number of variables randomly sampled as candidates at each decision node. Using 10 fold cross validation the optimum value was found to be 4 for Derby, 6 for Leicester and 8 for Nottingham. 1000 trees were found to produce a stable output for each model. The RF models for the three cities give a much better fit to the data than the GLS models explaining 98%, 97% and 96% of the variance in the crime deprivation data for Derby, Leicester and Nottingham respectively.

2.5.2.1 Feature selection

The next stage is to check whether each of the predictor variables included in the model have a significant effect on the CD dependant variable. The “Boruta” package in the R programming language (Kursa and Witold, 2010) provides a method for ranking the relative importance of each of the predictor variables and whether they are

statistically significant. The method works by firstly, adding randomness to the data set by creating shuffled copies of all predictor variables (which are called shadow features). Then, it trains the RF model on the extended data set and applies a feature importance measure (the mean decrease in accuracy when a variable is excluded) to evaluate the importance of each feature where higher means are more important. After each iteration, it checks whether a real variable has a higher importance than the best of its shadow features (i.e. whether the feature has a higher Z score than the maximum Z score of its shadow features) and removes features which are deemed highly unimportant. Finally, the algorithm stops either when all features gets confirmed or rejected or it reaches a specified limit of RF runs. The outputs from the Boruta algorithm for each of the RF models are shown in Figure 12 for Derby and Figures S13 and S14 in the SI for Leicester and Nottingham. For Leicester and Nottingham, the three soil parameters have the lowest importance with Ce being considered not significant for the Leicester model. Whilst the Boruta algorithm suggest that both Pb and Sn are significant predictors their median importance lies below the upper whisker of the “shadowMax” importance suggesting their effect influence on the model is marginal. This is in line with the GLS models findings.

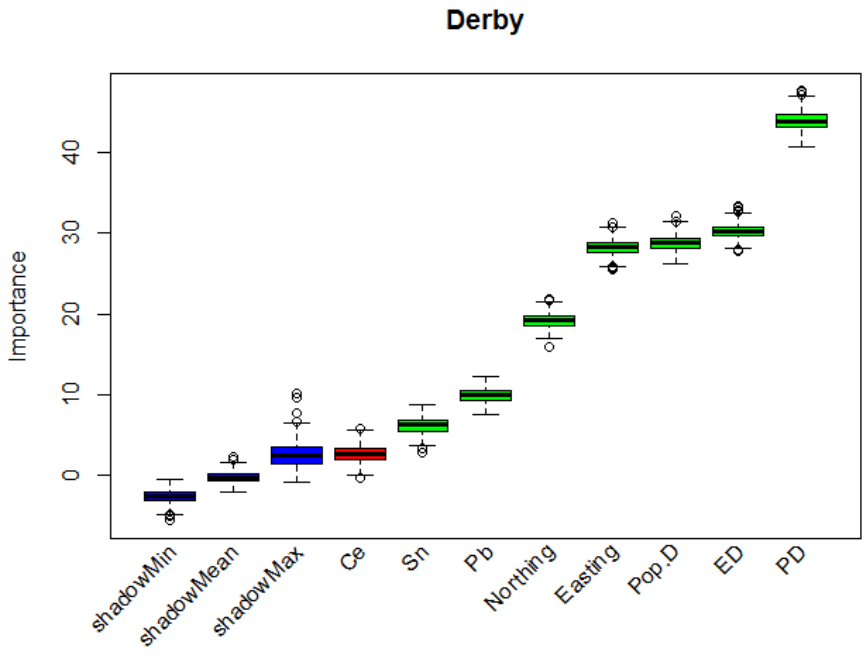

Figure 12 Feature selection output from the Boruta algorithm for Derby. Green represents significant features, red features are not significant and blue features represent the shadow features

The Derby model (Figure 12), however, shows the Pb mean importance to be above the shadowMax upper whisker backing up the findings for the GLS model that Pb in soil has a significant effect on crime deprivation in Derby.

2.5.2.2 Sensitivity Check

Examining the sensitivity of an RF model to a particular predictor variable, in this case Pb, is not straight forward as it depends not only on the value of predictor variable but also on the values set for the other co-variates. One approach (Goldstein et al., 2015) and its associated R programming language package “ICEbox” suggests plotting a family of partial dependence plots in which the values of the covariate conditions for each case used in the training the model are used and the Pb is varied over the range of values used in the training data. This produces a set of partial curves which

represent the behaviour of Pb under all the covariate conditions in the training set and gives an overview of the different ways in which Pb can affect the crime deprivation in Derby. Given the RF model and the original training data the ICEbox package within R calculates each partial curve for Pb and uses k-means clustering to group together partial curves with a similar shape. In this instance 3 groupings were found to give suitable clusters. Figure 13 shows the median partial curves for each of the identified cluster centred on the same starting location to show the relative absolute effect. Eleven percent of the partial curves (cluster 3) have a sigmoidal median partial curve showing an increase in transformed CD of 0.45 units between -1 to 1 transformed Pb in soil concentration units. The partial curves associated with cluster 2 (46% of the data) has a similar sigmoidal shape to cluster 1 with a lower increase in transformed CD of ca.2 units.

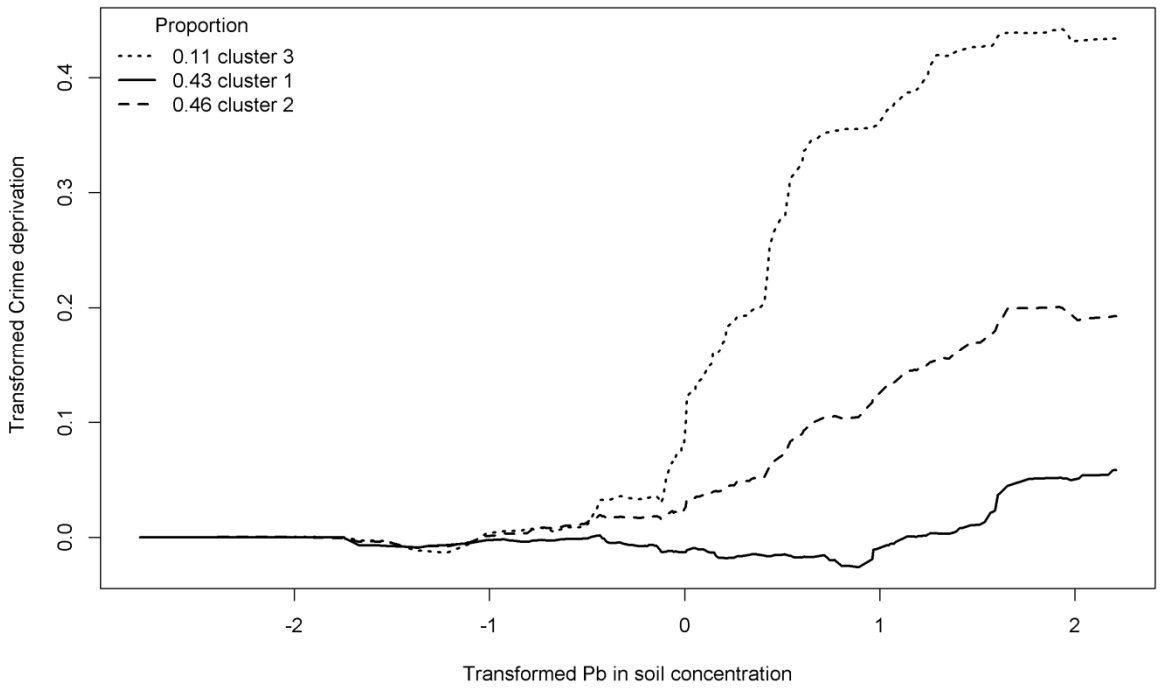

Figure 13 Median partial curves of the identified clusters for the effect of Pb in soil on crime deprivation in Derby showing the proportion of the curves associated with each cluster

Cluster 1 has a much flatter response than clusters 2 and 3 with a maximum of ca.0.5 transformed CD units occurring at higher values of transformed Pb in soil concentration units than clusters 2 and 3 (i.e. 1-2 units).

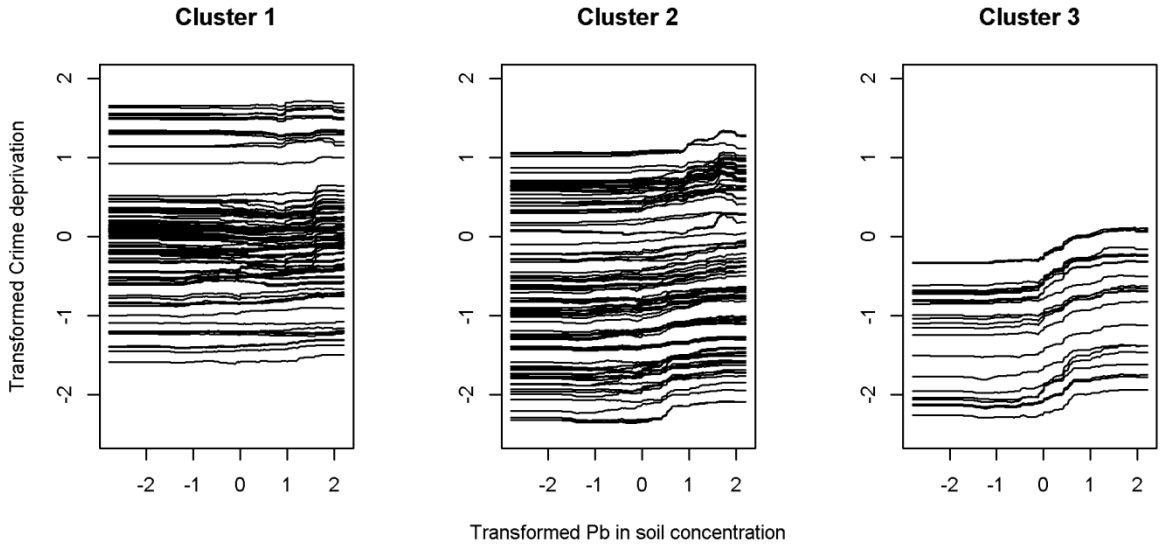

Figure 14 Individual partial Pb profiles associated with each cluster

Figure 14 shows the individual partial Pb profiles for Derby (without centering) associated with each of the clusters and shows that there is a general trend of decreasing CD in the order cluster1>cluster2>cluster3.

Figure 15 shows the median values of the socio-economic predictor variables and Pb in the three clusters. The personal deprivation (PD) predictor variable shows the greatest contrast between the three clusters with decreasing values in the order cluster 1>cluster 2>cluster3. The RF model shows that the greatest effect of Pb on CD occurs at lower PD (see cluster 3 in Figure 13and Figure 14) and as PD increases (cluster2 to cluster1) the effect of Pb in soil on CD decreases until in cluster 1 the effect is much reduced and only occurs at higher Pb concentrations. This clearly shows an antagonistic interaction effect between Pb in soil and PD. In addition to the interaction between Pb and PD, the shape of the partial profiles

provides some insight into the mechanism whereby Pb interacts with the human subject to cause higher crime rates. In Figure 13 the median curves for clusters 2 and 3 have a similar shape where, initially, an increase in Pb has no effect on CD until it reaches a trigger point concentration (ca.-0.5 transformed Pb units) where an increase in Pb causes a relatively sharp rise in CD until it reaches a plateau (ca. 1.7 transformed Pb units) where increasing Pb concentrations has little or no effect on CD. This sigmoidal shape is very much in agreement with biological studies in which the effect of increasing Pb concentrations on measured blood parameters shows a similar sigmoidal shape (Zielhuis, 1975). This provides some supporting evidence to our original hypotheses that it is the effect of Pb on human development that eventually leads to increased crime.

2.5.2.3 RF model uncertainty

In order to test whether the effect of Pb in soil on CD, as indicated in Figure 13 and Figure 14, are larger than the uncertainties in the RF modelling process the uncertainty needs to be quantified. Since the RF modelling process randomly resamples the data set every time it is run then running the model a number of times produces a simple bootstrap uncertainty estimate on the model output. Figure 16 shows median example profiles from each of the three clusters along with 95th percentile uncertainty limits calculated by running the model 500 times for each of the 3 sets of conditions. The upper confidence limit for the lowest Pb concentration for each example is marked as a horizontal line and, in all three instances, the predicted CD at high Pb, taking into account the confidence interval, exceeds the initial CD values. This clearly shows that the effect of Pb on CD in Derby is larger than the model uncertainty and helps to confirm the findings of the Boruta analysis (Figure 12) that Pb is an important predictor

variable for crime deprivation.

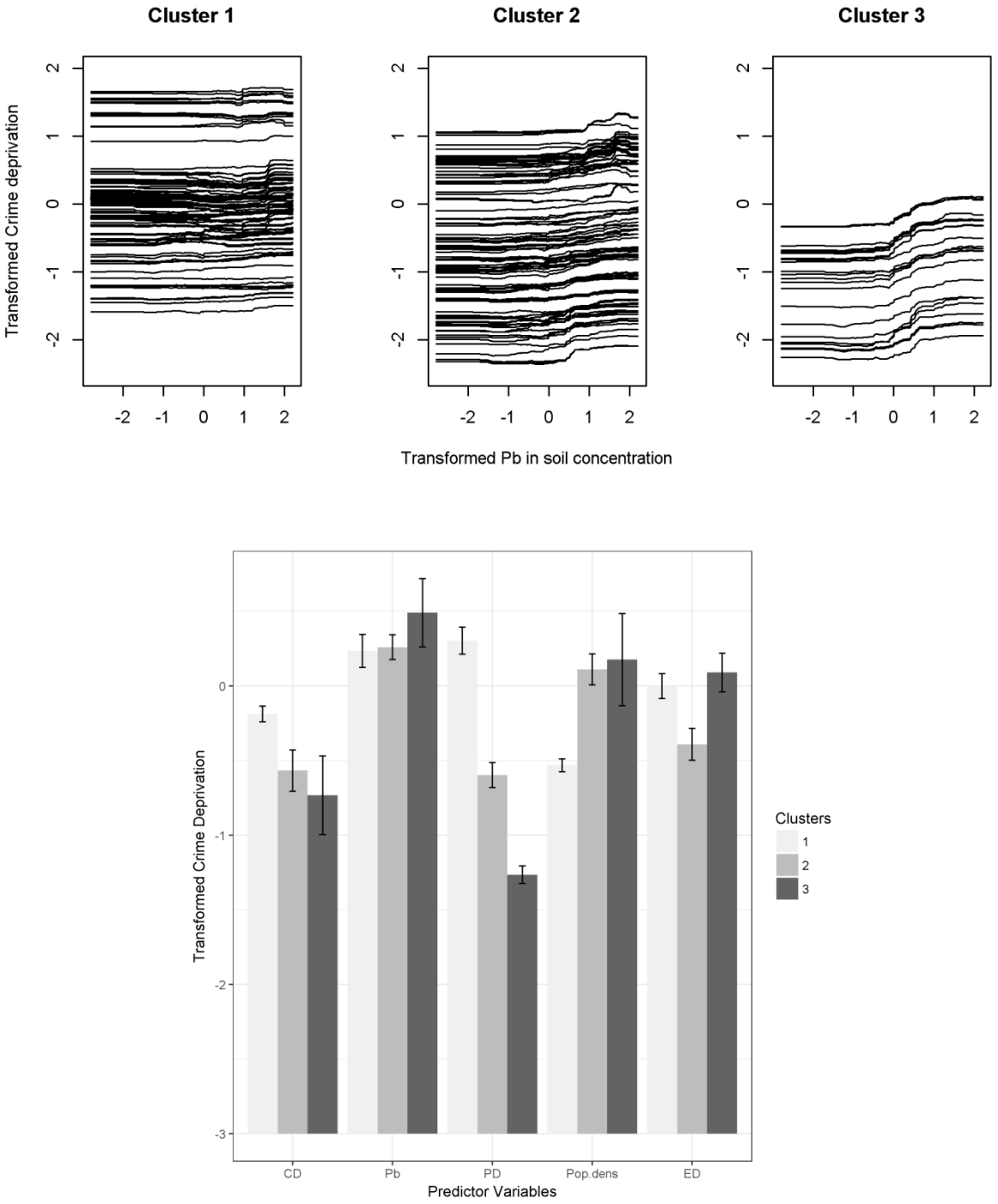

Figure 15 Median values associated with selected variables associated with each cluster for Derby. Error bars represent the standard error. CD is crime deprivation, PD is personal deprivation, Pop.dens is population density, and ED is environmental deprivation.

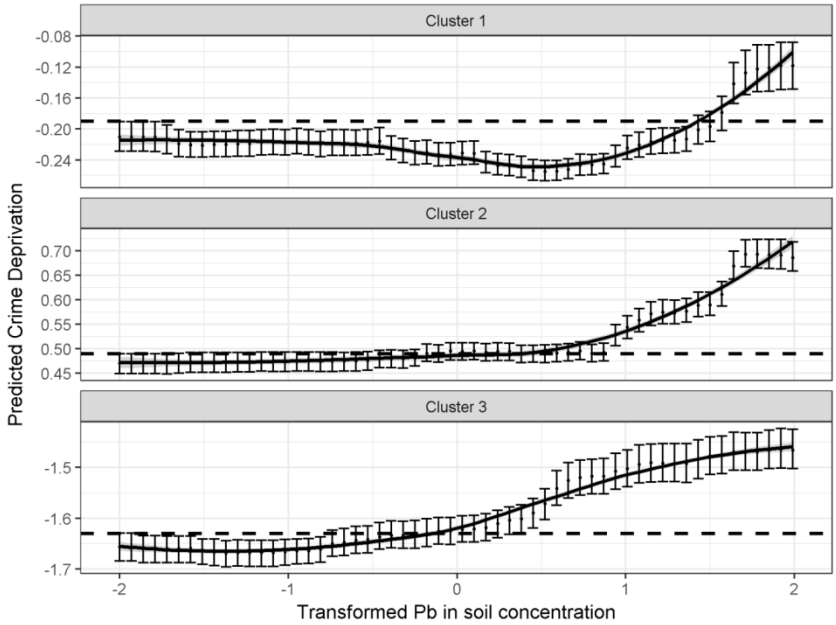

Figure 16 Example Pb partial profiles from each of the three clusters for Derby. The dotted horizontal line is the upper confidence limit at the lowest Pb concentration

2.5.2.4 Relative Effect of Pb on Crime deprivation in Derby

Whilst Figure 16 shows that increasing Pb concentrations in soil gives a significant increase in crime deprivation, the absolute size of the effect compared to the variation in crime deprivation as measured over all of Derby provides a measure of the overall importance of Pb in soil in controlling crime deprivation.

Figure 17 summarises, in a boxplot broken down by cluster, the difference between the maximum and minimum values of each of the partial Pb profiles, shown in Figure 14, ratioed to the interquartile range of the crime deprivation scores for the whole of Derby.

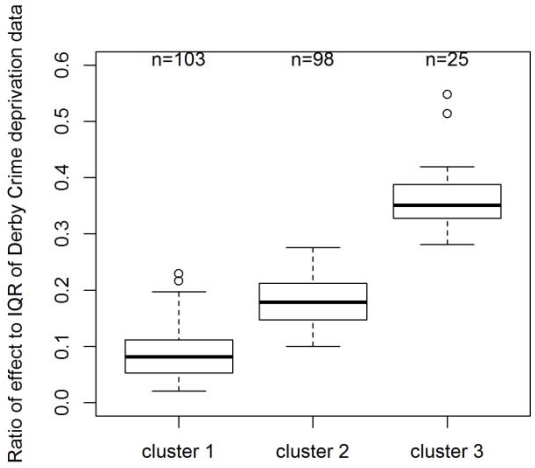

Figure 17 Ratio of the effect of increasing Pb in soil concentration to the interquartile range of the overall crime deprivation data for Derby

The more commonly found Pb partial profiles associated with clusters 1 and 2 have ratios ranging from about ca.0.02 to 0.28 whereas cluster 3 partial plots, which are less prevalent and occur at low PD (Figure 15, cluster 3), ranges from ca.0.3 to 0.4. Whilst the biggest effects are observed in areas of lower PD the effect of Pb on CD in all three cluster groupings is far from being negligible in the context of the range of CD values found over the whole of Derby.

2.5.2.5 Interpretation of results

Bringing together all the evidence from the different data sources (geological, geochemical, socio-economic and statistical data modelling) we can provide a holistic overview the relationship between CD in the three cities and the soil Pb concentrations.

From the geology we can see the spatial distribution of Pb in Derby follows the river flood plain sediment carried down from the Peak District. Studies have shown elevated

Pb concentrations in the stream sediments of the Derwent catchment (Kossoff et al., 2016) arising from the naturally elevated Pb in the soils and mine spoil from Pb mining. Pb in Leicester and Nottingham, however, show no clear association with geology more association with central regions of cities i.e. associated with anthropogenic sources. This provides initial evidence as to why the Pb content of the soils in Derby is higher than in Leicester and Nottingham (Figure 5) and also that the Pb in in the soil in Derby is likely to be in a different physico-chemical form than that in the other two cities. If we now consider the geochemistry of the soils and look at the highest spearman correlations of Pb with other elements in the three cities (Figure 18) we can gain further insight into the source of Pb. We find that in Leicester and Nottingham Pb is most highly correlated with Cu, Sb, Sn and Zn; all elements which are associated with anthropogenic inputs (Albanese and Breward, 2011). For Derby, however, Pb is most highly associated with Zn, Cd, Mo, Ba and Sr. A study of the superficial deposits in Peak District (Burek and Cubitt, 1979) establishes a link between Ca, Ba, Sr and Mo as representing typical chemical relationships developed in carbonate rocks in Derbyshire which supports our findings from the geological evidence.

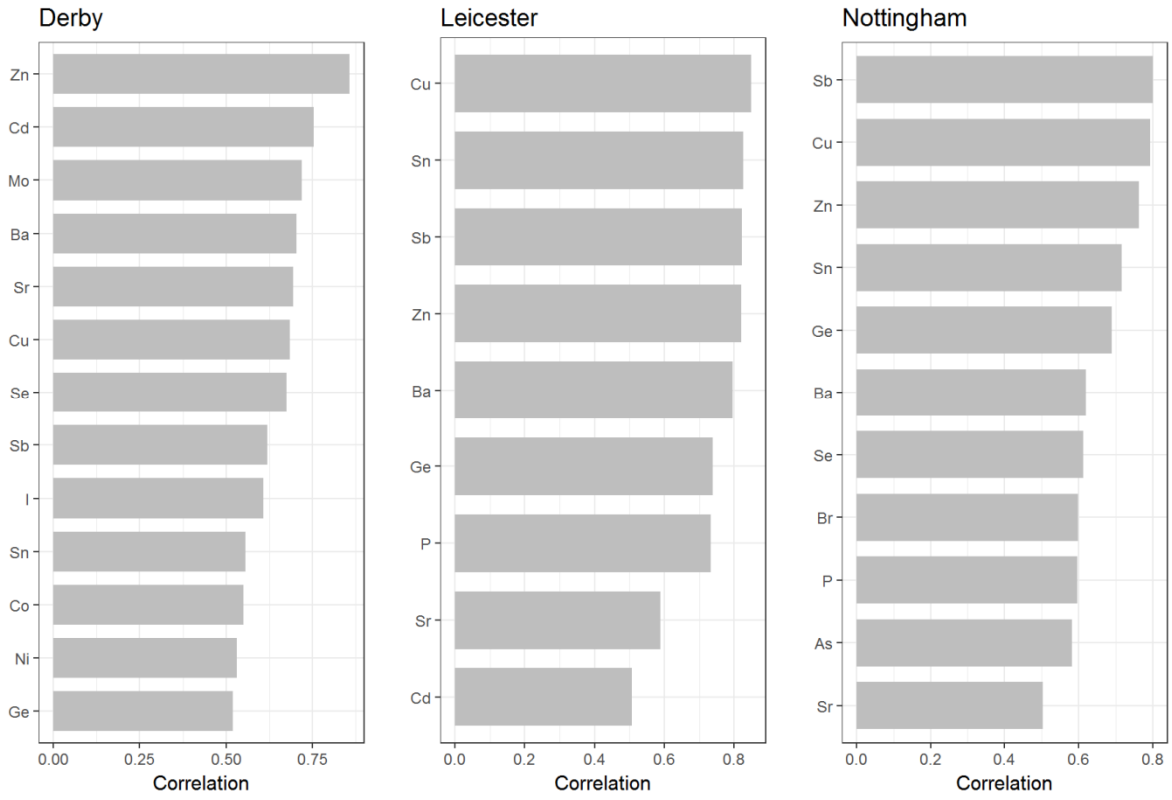

Figure 18 Comparison of the highest Spearman correlation coefficients (>0.5) of Pb with other elements in the soils of Derby, Leicester and Nottingham

The statistical modelling clearly shows that Pb has a significant effect on CD in Derby but not in Leicester or Nottingham. The effect on CD in Derby is of similar magnitude to the overall variability of CD in Derby and the shape of the response curve is in agreement with Pb dose response curves for biological studies of the effect of Pb on biochemical and haematological parameters in blood (Zielhuis, 1975) The RF model outputs can be used to provide some guidance on the concentrations of Pb in soil in Derby that start to have an effect on the human population. The average Pb partial profile curves for the three different clusters (Figure 13) can be considered as dose response curves. These have been re-plotted in Figure 19 with the Pb values converted back to the original concentration in the soil. For clusters 1 and 2 the lowest

observable effect level (LOEL) is ca.100-150 mg/kg Pb. For cluster 3 which is associated with higher PD the LOEL is much higher at ca.300-500mg/kg.

Given the RF model outputs for CD in Derby, explanations for the reasons why Pb in soil in Leicester and Nottingham was not found to be an important predictor of CD can postulated. Firstly, Pb in soil in Derby was found to be higher than in Leicester and Nottingham (Figure 5) so there are less soil samples above the LOEL of 150mg/kg for Leicester and Nottingham (150 mg/kg is 69th percentile in Leicester,61st percentile in Nottingham and 47th percentile in Derby).

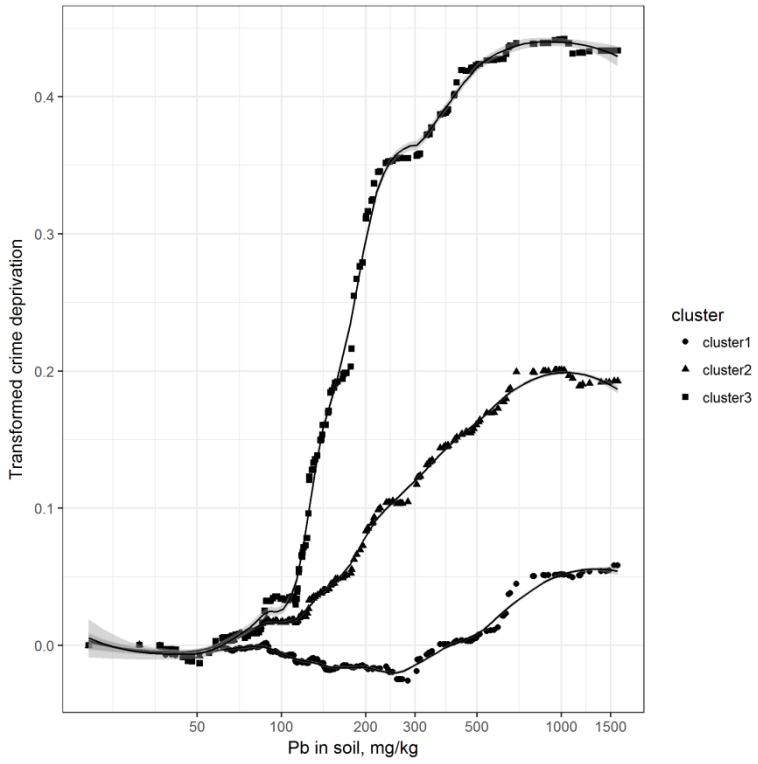

Figure 19 Mean dose response curves for the three response clusters for Derby

Both Leicester and Nottingham have higher CD (Figure 7) and PD (not shown) and the RF model for Derby shows that higher PD reduces the effect of Pb as a predictor variable (compare cluster 1 with cluster 3 in Figure 19).

In addition to this, the geological and geochemical evidence clearly shows that Pb in Derby comes from a different source than Leicester or Nottingham and is therefore likely to be in a different physico-chemical form in the soils and therefore its mobility and bioavailability will be different (Appleton et al., 2013). . If the hypotheses that Pb is causing detrimental effects to neurological development in children there must be a pathway for the Pb in the soil to enter the human body. Risk assessment studies clearly show that direct ingestion of soil and inhalation of soils dusts (Swartjes, 2011) are important routes for metal uptake into the human body from soils. This provides the final link in testing the hypotheses that Pb in soil has an effect on crime rate which, given the evidence found in this study, is shown to be true under certain conditions. The major controlling features are the concentration and form of Pb in the soil and the level of socio economic deprivation in the city under study.

3 Conclusion

Geological and geochemical evidence on the soil sampling sites and the composition of the soils under study provides important information to help understand the link between soil geochemistry and socio-economic factors of the human population. Both the GLS and the RF statistical models showed that the socio-economic predictor variables (particularly PD) and spatial associations were important in predicting crime deprivation. Pb and the two other soil chemistry parameters (Sn and Ce) were not important predictors of crime deprivation in Leicester and Nottingham. Pb and its interactions with spatial and socio-economic factors were, however, shown to have a significant effect on crime deprivation in Derby.

Interrogation of the RF model for Derby showed that there was an antagonistic interaction effect between Pb in soil and PD. Where there is low personal deprivation,

Pb in soil has an effect that can be up half of the interquartile range of crime deprivation variation found over all of Derby (Figure 17). The effect of Pb on CD reduces and the effect of PD on CD takes over as the controlling factor at high PD values. The RF model can be used to produce “dose-response” curves of the effect of Pb in soil on crime deprivation under different spatial and socio-economic conditions.

These findings may have implications going forward for risk assessment methodologies which assume that the dose response curves for contaminants are constant and independent of spatial and socio-economic factors which this study suggests is not always the case.

References

Aide, M.T., Aide, C., 2012. Rare Earth Elements: Their Importance in Understanding Soil Genesis. ISRN Soil Science 2012, 11.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control 19, 716-723.

Albanese, S., Breward, N., 2011. Sources of anthropogenic contaminants in the urban environment. Mapping the chemical environment of urban areas. Wiley-Blackwell, Oxford, 116-127.

Allen, M.A., Cave, M.R., Chenery, S.R.N., Gowing, C.J.B., Reeder, S., 2011. Sample Preparation and Inorganic Analysis for Urban Geochemical Survey Soil and Sediment Samples, Mapping the Chemical Environment of Urban Areas. John Wiley & Sons, Ltd, pp. 28-46.

Ander, E.L., Johnson, C.C., Cave, M.R., Palumbo-Roe, B., Nathanail, C.P., Lark, R.M., 2013. Methodology for the determination of normal background concentrations of contaminants in English soil. Science of the Total Environment 454–455, 604-618.

Appleton, J.D., Cave, M.R., Palumbo-Roe, B., Wragg, J., 2013. Lead bioaccessibility in topsoils from lead mineralisation and urban domains, UK. Environmental Pollution 178, 278-287.

ATSDR 2007. Toxicological Profile for Lead. Agency for Toxic Substances and Disease Registry, Atlanta.

Bellinger, D.C., 2008. Lead neurotoxicity and socioeconomic status: Conceptual and analytical issues. NeuroToxicology 29, 828-832.

Box, G.E.P., Draper , N.R., 1987. Empirical Model-Building and Response Surfaces. Wiley.

Breiman, L., 2001a. Bagging Predictors. Machine Learning 24, 123-140.

Breiman, L., 2001b. Random Forests. Machine Learning 45, 5-32.

Burek, C.V., Cubitt, J.M., 1979. Trace element distribution in the superficial deposits of Northern Derbyshire, England. Minerals and the Environment 1, 90-100.

Flight, D., Scheib, A.J., 2011. Soil Geochemical Baselines in UK Urban Centres: The G-BASE Project. Wiley Online Library.

Goldstein, A., Kapelner, A., Bleich, J., Pitkin, E., 2015. Peeking Inside the Black Box: Visualizing Statistical Learning With Plots of Individual Conditional Expectation. Journal of Computational and Graphical Statistics 24, 44-65.

Johnson, C.C., 2011. Understanding the Quality of Chemical Data from the Urban Environment – Part 1: Quality Control Procedures, in: Johnson, C.C., Demetriades, A., Locutura, J., Ottesen, R.T. (Eds.), Mapping the Chemical Environment of Urban Areas. Wiley-Blackwell, Oxford, pp. 61-76.

Kossoff, D., Hudson-Edwards, K.A., Howard, A.J., Knight, D., 2016. Industrial mining heritage and the legacy of environmental pollution in the Derbyshire Derwent catchment: Quantifying contamination at a regional scale and developing integrated

strategies for management of the wider historic environment. Journal of Archaeological Science: Reports 6, 190-199.

Kuhn, M., 2016. caret: Classification and Regression Training, R package version 6.0-71 ed.

Kursa, M.B., Witold, R.R., 2010. Feature Selection with the Boruta Package. Journal of Statistical Software 36, 1-13.

Laidlaw, M.A.S., Taylor, M.P., 2011. Potential for childhood lead poisoning in the inner cities of Australia due to exposure to lead in soil dust. Environmental Pollution 159, 1-9.

Liaw, A., Wiene, M., 2002. Classification and Regression by randomForest. R News 2, 18-22.

Marcus, D.K., Fulton, J.J., Clarke, E.J., 2010. Lead and Conduct Problems: A Meta-Analysis. Journal of Clinical Child & Adolescent Psychology 39, 234-241.

Mielke, H.W., Zahran, S., 2012. The urban rise and fall of air lead (Pb) and the latent surge and retreat of societal violence. Environment International 43, 48-55.

Nevin, R., 2007. Understanding international crime trends: The legacy of preschool lead exposure. Environmental Research 104, 315-336.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., R Core Team, 2016. nlme: Linear and Nonlinear Mixed Effects Models, package version 3.1-128 ed.

R Core Team, 2016. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.

Rawlins, B.G., McGrath, S.P., Scheib, A.J., Breward, N., Cave, M., Lister, T.R., Ingham, M., Gowing, C., Carter, S., 2012. The Advanced Soil Geochemical Atlas of England and Wales. . British Geological Survey, Nottingham

Scheib, A.J., Nice, S.E., 2008. Soil geochemical baseline data for the urban areas of Corby, Coventry, Derby, Leicester, Northampton, Nottingham and Peterborough in the East Midlands. British Geological Survey.

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., Plunkett, E., 2015. The English Indices of Deprivation 2015. Department for Communities and Local Government.

Swartjes, F.A., 2011. Dealing with contaminated sites: from theory towards practical application. Springer Science & Business Media.

Taylor, M.P., Forbes, M.K., Opeskin, B., Parr, N., Lanphear, B.P., 2016. The relationship between atmospheric lead emissions and aggressive crime: an ecological study. Environmental Health 15, 23.

Vuong, Q.H., 1989. Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses. Econometrica 57, 307-333.

Wikström, P.O., Oberwittler, D., Treiber, K., Hardie, B., 2012. Breaking rules: The social and situational dynamics of young people's urban crime. OUP Oxford.

Wright, J.P., Dietrich, K.N., Ris MD, Hornung, R.W., Wessel, S.D., Lanphear, B.P., Ho, M., Rae, M.N., 2008. Association of Prenatal and Childhood Blood Lead Concentrations with Criminal Arrests in Early Adulthood. PLoS Med 5.

Yeo, I.K., Johnson, R.A., 2000. A new family of power transformations to improve normality or symmetry. Biometrika 87, 954-959.

Young, A., Bryant, E., Winchester, H., 1992. The Wollongong lead study: an investigation of the blood lead levels of pre-school children and their relationship to soil lead levels. Aust Geogr 23.

Zahran, S., Mielke, H.W., Gonzales, C.R., Powell, E.T., Weiler, S., 2010. New Orleans before and after Hurricanes Katrina/Rita: A Quasi-Experiment of the Association

between Soil Lead and Children's Blood Lead. Environmental Science & Technology 44, 4433-4440.

Zielhuis, R.L., 1975. Dose-response relationships for inorganic lead. International Archives of Occupational and Environmental Health 35, 19-35.

Zuur, A.F., Ieno, E.N., Saveliev, A.A., 2009. Mixed Effects Models and Extensions in Ecology with R. Springer.