This thesis has been submitted in fulfilment of the requirements for a postgraduate degree (e.g. PhD, MPhil, DClinPsychol) at the University of Edinburgh. Please note the following terms and conditions of use:

This work is protected by copyright and other intellectual property rights, which are retained by the thesis author, unless otherwise stated.

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge.

This thesis cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author.

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author.

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given.
Predictive modelling of soil lead in urban environments.

Sarah L. Donoghue

Doctor of Philosophy
The University of Edinburgh
2022
Declaration

I declare that this thesis has been composed solely by myself and that it has not been submitted, in whole or in part, in any previous application for a degree or professional qualification. Except where states otherwise by reference or acknowledgment, the work presented is entirely my own.

Sarah Donoghue

October 2022
Acknowledgements

Firstly, I would like to thank my funders for making this PhD possible and for the training they provided. They are the Edinburgh Earth and Environment (E³) Doctoral Training Partnership (DTP) funded by the National Environment and Research Council (NERC), the University of Edinburgh, and the British Geological Survey (BGS). The British Geological Survey also kindly provided me with the data which I needed to complete this project.

I would also like to thank my supervisory team for their invaluable help and support. This includes Prof. Margret Graham for leading the project and for her advice on soil geochemistry, Dr Neil Stuart for his insightful comments and knowledge on spatial analysis, Dr Benjamin Marchant for his patience whilst explaining the intricacies of geostatistics and coding in Matlab, Fiona Fordyce for her practical counsel on soil sampling and analysis, and the workings of BGS, and lastly Prof. Murray Lark for introducing spatial statistics at the start of the project. Additionally, thank you to Dr Lorna Eades for her assistance in running the ICP-OES/MS analysis, and all my lab colleagues for their help and advice both in the lab and the field - Gavin Sim, Andrew Innes, Sam Carter, Michael Muir, and James Watt.

Lastly, I thank my family – especially Christine and Tina Beadle for their many hours of amazing help and enthusiasm while collecting field samples, sometimes in challenging conditions. My husband, Christopher Donoghue, for giving me endless support and patience while I completed this PhD, even though it took a lot longer than first intended. And last but definitely not least, my children (and everyone who looked after them) - Lily, Alexander, and Albert Donoghue – for giving me the time and space to work on my own project.

To everyone else who believed in me and supported me through this PhD, thank you.
Abstract

Lead (Pb) is a naturally occurring, potentially toxic, element, which may result in health effects. In particular, Pb is a neurotoxin so can cause neurological damage (for example lower IQ scores) even at low exposure levels. Young children, under six years old, are at particular risk. An important possible exposure pathway is through ingestion or inhalation of contaminated urban soil. Locating soils with high Pb concentrations traditionally involves costly and time-consuming sampling, laboratory analysis, and mapping. Machine learning techniques that utilise existing data from similar urban areas offers a potentially valuable short-cut.

This research analyses three soil Pb datasets from varying scales in Greater Glasgow; the city scale British Geological Survey's (BGS) Geochemical Baseline of the Environment (G-BASE) dataset, and two new neighbourhood scale datasets from Paisley and Bishopbriggs. All samples were collected using a non-probabilistic, grid sampling strategy. Consequently, spatial dependence is present in all datasets and model-based analysis is required to account for spatial error. Spatial variability in each dataset was assessed and mapped using [local] Moran's I analysis, and also restricted maximum likelihood (REML) fit semivariograms with ordinary kriging. Hotspots and coldspots of soil Pb concentrations were visible in all study areas.

Spatial heterogeneity in soil Pb concentrations were explored by examining the relationship between soil Pb concentrations and possible covariates. Potential covariates were selected from the literature and collated in GIS using nearest neighbour joins. The relationship between each covariate and soil Pb concentrations was assessed using model-based geostatistics i.e. linear mixed models (LMM) and Wald tests. Additionally, each covariate was included when mapping soil Pb concentrations using the empirical best linear unbiased predictor (E-BLUP). Overall, only covariates which showed an almost significant ($p < 0.06$) and justifiable relationship with soil Pb concentrations were input in machine learning models. However, geographical scale and location can impact covariate influence on soil Pb.

In the Glasgow city dataset six covariates (soil organic matter (OM) percentage, land use, building age, and historic industry type, age, and density) were used to train a random forest (RF) model. This was combined with quantile regression forests (QRF) to predict soil Pb concentrations for a range of percentiles and for
different soil Pb groups i.e. only high samples containing ≥200 mg kg⁻¹ of soil Pb. The accuracy of these predictions was determined using 10-fold cross validation and was reasonably good; the concordance correlation coefficient (CCC) for all samples at the 50th percentile is 0.33, and the Pearson’s R² value is 0.51. Higher percentiles (i.e. the 75th or 95th percentile) more accurately predicted samples containing high Pb concentrations. This knowledge could be used to more accurately locate soil Pb hotspots.

QRF was also used to predict areas of the city at higher risk (≥30% probability) of containing high soil Pb concentrations (≥200 mg kg⁻¹). This probability was mapped for a prediction grid covering the city to give a city-wide map of soil Pb risk, with hotspots from Getis-Ord Gi∗ cluster analysis added as contour lines. The prediction accuracy was assessed by comparing measured and predicted soil Pb groups using a confusion matrix, where the overall accuracy was good at 74% and the high user’s/producer’s accuracies were >50%.

In the neighbourhood datasets different covariates were used to train a RF model; soil OM, land use, building age, building distance, and road distance in Paisley and soil OM, land use, historic industry distance and type in Bishopbriggs. Validation statistics indicated that the models were reasonably accurate, i.e. the CCC for all samples at the 50th percentile was 0.35 and 0.56 for Bishopbriggs and Paisley respectively (Pearson’s R² = 0.59 and 0.62). Although the Paisley model had better overall accuracy, the Bishopbriggs model was more accurate at predicting sample with high soil Pb concentrations using the 75th percentile (CCC = 0.75). The Paisley and Bishopbriggs RF models were also used with QRF to predict the probability of an unknown location containing high soil Pb concentrations. These probability predictions showed very good overall accuracy (83% and 84% for Bishopbriggs and Paisley respectively), with 98% producer’s accuracy for high samples in Paisley.

This research is unique as it tests the accuracy of a RF model in other cities which it was not trained in. When the Glasgow QRF model was used to predict locations at high-risk from high soil Pb concentrations in Belfast and Leicester, there was a 71/80% overall accuracy and 53/47% producer’s accuracy to known high Pb soil samples respectively. Consequently, the machine learning model can help locate the areas of a city most at-risk from high soil Pb concentrations to target subsequent soil sampling and eventual remediation.
Lay abstract

Lead is a naturally occurring metal that has many useful properties. During the 19th and 20th centuries it was used in industry, petrol, paints, and pipes. This meant that large amounts of lead entered the soil, especially in cities. This lead can remain in the soil for hundreds of years, during which time it can potentially impact our health. In particular, lead can damage brain development, especially in young children under the age of six. There is no safe level of lead in the blood, and even low levels of blood lead can result in a reduced intelligence quotient (IQ) and other health impacts.

In order to reduce the potential health impacts from soil lead, we first need to know where high levels of soil lead can be found. This requires a map. Traditionally, soil lead maps are made by going to a city, collecting hundreds or thousands of soil samples, and then analysing how much lead is contained in each soil sample in the laboratory. Computer modelling can then use these lead values at known sample points to predict soil lead values for a whole city. However, this usually requires a lot of time and money.

Machine learning is an alternative method to generate a map of soil lead for an entire city. It relies on the relationship between soil lead and other factors, for example the distance to a road, the land use at the site, and the age of the nearest house. Random forests are a type of machine learning method which use a training dataset to build multiple decision trees. Each tree grows from a random subsample of the training data, also known as the root. Branches then split off the root data based on criteria, for example the distance to the nearest historic industry at each sample point or the organic matter content in the soil. At the end of some branches are leaves, where soil lead values are averaged for samples meeting all the branch criteria. These trees can then use other datasets e.g. historic industry distances and soil organic matter contents, to predict soil lead values in a city that has not been sampled.

This study built random forest models using soil lead values in Glasgow. One random forest model was built at the city scale, using data collected by the British Geological Survey. Two other random forest models were constructed at the neighbourhood scale, using data I collected from Paisley and Bishopbriggs.
Careful analysis was first run to decide which factors should predict soil lead in each model. The choice of factors varied at different locations and scales.

The random forest models were first tested in the area they were trained in (i.e. Glasgow). The models were used to predict which collected samples had a high chance (at least a 30% probability) of containing high soil lead levels (at least 200 mg kg$^{-1}$). This was then compared with samples which actually had high lead levels (at least 200 mg kg$^{-1}$). The models performed pretty well, as the Glasgow city scale model had a 74% success rate, the Paisley and Bishopbriggs models had 83-84% success rates.

The models were also tested on other British cities – Belfast and Leicester – that they had not been trained in. The predictions were then compared with known soil lead values in these cities that had already been measured by the British Geological Survey. Again, the models were able to predict high risk locations reasonably well, as the success rate was 71% and 80% in Belfast and Leicester.

The random forest models have several practical uses. For example, a local authority could use the city scale model to pinpoint the areas of a city most at risk from high soil lead values. Alternatively, a homeowner could use the neighbourhood model to decide if they should investigate soil lead contents in their garden soil. Overall, it is hoped that random forest models, like those developed in this study, will be included as part of land contamination assessments, which decide if further investigation of a site is required before new development.
Abbreviations

AIC  Akaike information criterion
ANN  Artificial neural networks
AUC  Area under the curve
BGS  British Geological Survey
BNG  British national grid
C4SL Category 4 screening level
CART Classification and regression trees
CCC  Concordance correlation coefficient
CDC  Centres for Disease Control and Prevention
CLEA Contaminated Land Exposure Assessment
CoDA Compositional data analysis
CRM  Certified reference materials
DEFRA Department for Environment, Food, and Rural Affairs
DOM  Dissolved organic matter
E-BLUP Empirical best linear unbiased predictor
EDI  Estimated daily intake
G-BASE Geochemical Baseline Survey of the Environment
GFAAS Graphite furnace atomic absorption spectroscopy
GLM  General linear model
GPS  Geographical positioning system
HHI  Heavy historical industry
HQ  Hazard quotient
ICP-MS Inductively coupled plasma mass spectrometry
ICP-OES Inductively coupled plasma optical emission spectroscopy
ICRCL Inter-departmental Committee for the Redevelopment of Contaminated Land
IDW  Inverse distance weighting
IQ  Intelligence quotient
IQR  Interquartile range
LCRM Land contamination risk assessment
LHI  Light historical industry
LISA Local indicators of spatial autocorrelation
LMM  Linear mixed model
LOD  Limit of detection
<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOI</td>
<td>Loss on ignition</td>
</tr>
<tr>
<td>LOOCV</td>
<td>Leave-one-out cross validation</td>
</tr>
<tr>
<td>ME</td>
<td>Mean error</td>
</tr>
<tr>
<td>Mean.AD</td>
<td>Mean absolute difference /deviation</td>
</tr>
<tr>
<td>Med.AD</td>
<td>Median absolute difference /deviation</td>
</tr>
<tr>
<td>MOM</td>
<td>Method of moment</td>
</tr>
<tr>
<td>NHI</td>
<td>Non-historical industry</td>
</tr>
<tr>
<td>OK</td>
<td>Ordinary kriging</td>
</tr>
<tr>
<td>OLS</td>
<td>Ordinary least squares</td>
</tr>
<tr>
<td>OM</td>
<td>Organic matter</td>
</tr>
<tr>
<td>OOB</td>
<td>Out-of-bag</td>
</tr>
<tr>
<td>OS</td>
<td>Ordnance survey</td>
</tr>
<tr>
<td>PDP/PIP</td>
<td>Predictor dependence/importance plots</td>
</tr>
<tr>
<td>PHE/PTE</td>
<td>Potentially harmful/toxic element(s)</td>
</tr>
<tr>
<td>ppm</td>
<td>Parts per million</td>
</tr>
<tr>
<td>QRF</td>
<td>Quantile regression forests</td>
</tr>
<tr>
<td>REML</td>
<td>Restricted maximum likelihood</td>
</tr>
<tr>
<td>RF</td>
<td>Random forests</td>
</tr>
<tr>
<td>RMSE</td>
<td>Root mean squared error</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver operating characteristic</td>
</tr>
<tr>
<td>RSD</td>
<td>Relative standard deviation</td>
</tr>
<tr>
<td>SD</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>SGV/SSV</td>
<td>Soil guideline/screening value</td>
</tr>
<tr>
<td>SIMD</td>
<td>Scottish Index of Multiple Deprivation</td>
</tr>
<tr>
<td>(S)KDE</td>
<td>(Spatial) kernel density estimation</td>
</tr>
<tr>
<td>SSE</td>
<td>Standardised squared error</td>
</tr>
<tr>
<td>SVM</td>
<td>Support vector machines</td>
</tr>
<tr>
<td>TDI</td>
<td>Total daily intake</td>
</tr>
<tr>
<td>USEPA</td>
<td>U.S. Environmental Protection Agency</td>
</tr>
<tr>
<td>VPD/H</td>
<td>Vehicles per day/hour</td>
</tr>
<tr>
<td>WLS</td>
<td>Weighted least squares</td>
</tr>
<tr>
<td>XRF</td>
<td>X-ray fluorescence spectrometry</td>
</tr>
<tr>
<td>YI</td>
<td>Youden’s index</td>
</tr>
</tbody>
</table>
**Chemical elements and compounds**

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Element</th>
</tr>
</thead>
<tbody>
<tr>
<td>As</td>
<td>Arsenic</td>
</tr>
<tr>
<td>Cd</td>
<td>Cadmium</td>
</tr>
<tr>
<td>Cr</td>
<td>Chromium</td>
</tr>
<tr>
<td>Cu</td>
<td>Copper</td>
</tr>
<tr>
<td>Fe</td>
<td>Iron</td>
</tr>
<tr>
<td>Hg</td>
<td>Mercury</td>
</tr>
<tr>
<td>Mn</td>
<td>Manganese</td>
</tr>
<tr>
<td>Na</td>
<td>Sodium</td>
</tr>
<tr>
<td>Ni</td>
<td>Nickle</td>
</tr>
<tr>
<td>Pb</td>
<td>Lead</td>
</tr>
<tr>
<td>S</td>
<td>Sulphur</td>
</tr>
<tr>
<td>Sb</td>
<td>Antimony</td>
</tr>
<tr>
<td>Si</td>
<td>Silica</td>
</tr>
<tr>
<td>Sn</td>
<td>Tin</td>
</tr>
<tr>
<td>Th</td>
<td>Thorium</td>
</tr>
<tr>
<td>U</td>
<td>Uranium</td>
</tr>
<tr>
<td>Zn</td>
<td>Zinc</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chemical Formula</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$2\text{PbCO}_3\cdot\text{Pb(OH)}_2$</td>
<td>White lead</td>
</tr>
<tr>
<td>$\text{CaCl}_2$</td>
<td>Calcium chloride</td>
</tr>
<tr>
<td>$\text{CrO}_4\text{Pb}$</td>
<td>Lead chromate</td>
</tr>
<tr>
<td>$\text{HF}$</td>
<td>Hydrofluoric acid</td>
</tr>
<tr>
<td>$\text{HNO}_3$</td>
<td>Nitric acid</td>
</tr>
<tr>
<td>$\text{Pb}_2\text{Sb}_2\text{O}_7$</td>
<td>Lead antimonite</td>
</tr>
<tr>
<td>$\text{Pb}_3\text{O}_4$</td>
<td>Red lead</td>
</tr>
<tr>
<td>$\text{Pb}_5\text{OH(AsO}_4)_3$</td>
<td>Lead arsenate</td>
</tr>
<tr>
<td>$\text{PbO}$</td>
<td>Litharge</td>
</tr>
</tbody>
</table>
Contents

Chapter 1. Introduction .............................................. 1
  1.1. Project background ........................................... 1
  1.2. Research questions .......................................... 2
  1.3. Thesis outline ............................................... 2

Chapter 2. Literature review on urban soil Pb ...................... 5
  2.1. Health implications associated with Pb ...................... 5
  2.2. Soil ingestion/inhalation as an exposure pathway .......... 7
  2.3. How much soil Pb is too much? ............................ 11
  2.4. The importance of studying Pb in urban environments .... 13
  2.5. Why soil Pb is still a current issue ......................... 17
  2.6. The compositional nature of geochemical data ............. 18
  2.7. Conclusions .................................................. 19

References – Chapters 1 and 2 ........................................ 20

Chapter 3. Measuring total soil Pb concentrations ............. 29
  3.1. City scale measurements ..................................... 29
  3.2. Neighbourhood scale measurements ......................... 32
    3.2.1. Study area selection .................................... 32
      3.2.1.1. Paisley ................................................ 33
      3.2.1.2. Bishopbriggs ........................................ 35
    3.2.2. Sampling methodology .................................. 37
    3.2.3. Initial laboratory analysis ............................. 40
      3.2.3.1. Total Pb concentration ............................ 40
      3.2.3.2. pH .................................................... 42
      3.2.3.3. Organic matter .................................... 43
      3.2.3.4. Quality control checks ............................. 43

References – Chapter 3 ................................................ 50

Appendix 3.1 .......................................................... 52

Chapter 4. Exploratory data analysis .............................. 53
  4.1. Methodology .................................................. 53
  4.2. Chapter aims ................................................ 54
  4.3. Results and discussion ...................................... 55
    4.3.1. Soil Pb concentrations ................................. 55
    4.3.2. Skew ...................................................... 57
    4.3.3. Isotropy .................................................. 58

References – Chapter 4 ................................................ 61

Appendix 4.1 .......................................................... 62
<table>
<thead>
<tr>
<th>Chapter 7. Methodology</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.3. Methodology</td>
<td>193</td>
</tr>
<tr>
<td>7.4. Results and discussion</td>
<td>196</td>
</tr>
<tr>
<td>7.4.1. Quality control checks of isotopic results</td>
<td>196</td>
</tr>
<tr>
<td>7.4.2. Percentage and total Pb from petrol and coal sources</td>
<td>197</td>
</tr>
<tr>
<td>7.4.3. Modelling and mapping petrol and coal soil Pb</td>
<td>201</td>
</tr>
<tr>
<td>7.4.4. The influence of road distance on petrol and coal soil Pb</td>
<td>206</td>
</tr>
<tr>
<td>7.4.5. The influence of historic industry distance on petrol and coal soil Pb</td>
<td>211</td>
</tr>
<tr>
<td>7.5. Conclusions</td>
<td>215</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 8. Machine learning prediction of soil Pb at different scales</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.1. Background</td>
<td>219</td>
</tr>
<tr>
<td>8.1.1. Introduction to random forests</td>
<td>219</td>
</tr>
<tr>
<td>8.1.2. Studies using random forests to predict soil Pb</td>
<td>220</td>
</tr>
<tr>
<td>8.1.3. Defining high soil Pb values</td>
<td>220</td>
</tr>
<tr>
<td>8.2. Chapter aims</td>
<td>228</td>
</tr>
<tr>
<td>8.3. Methodology</td>
<td>228</td>
</tr>
<tr>
<td>8.4. Results and discussion</td>
<td>234</td>
</tr>
<tr>
<td>8.4.1. City scale predictions</td>
<td>234</td>
</tr>
<tr>
<td>8.4.1.1. Covariate selection</td>
<td>234</td>
</tr>
<tr>
<td>8.4.1.2. Prediction of soil Pb concentrations</td>
<td>235</td>
</tr>
<tr>
<td>8.4.1.3. Prediction of soil Pb risk</td>
<td>244</td>
</tr>
<tr>
<td>8.4.2. Neighbourhood scale predictions</td>
<td>249</td>
</tr>
<tr>
<td>8.4.2.1. Covariate selection</td>
<td>249</td>
</tr>
<tr>
<td>8.4.2.2. Prediction of soil Pb concentrations</td>
<td>252</td>
</tr>
<tr>
<td>8.4.2.3. Prediction of soil Pb risk</td>
<td>261</td>
</tr>
<tr>
<td>8.4.3. Practical uses</td>
<td>264</td>
</tr>
<tr>
<td>8.5. Conclusions</td>
<td>268</td>
</tr>
<tr>
<td>References - Chapter 8</td>
<td>270</td>
</tr>
<tr>
<td>Appendix 8.1 - Example random forest Matlab code</td>
<td>274</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 9. Modelling in other cities and geographical locations</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1. Introduction to new datasets</td>
<td>279</td>
</tr>
<tr>
<td>9.2. Chapter aims</td>
<td>284</td>
</tr>
<tr>
<td>9.3. Methodology</td>
<td>284</td>
</tr>
<tr>
<td>9.4. Results</td>
<td>285</td>
</tr>
<tr>
<td>9.4.1. Prediction of soil Pb concentrations - city scale</td>
<td>285</td>
</tr>
</tbody>
</table>
Figures

2.1 IQ vs blood Pb concentrations (From Canfield et al., 2003) .......... 6
2.2 Main Pb exposure pathways ........................................ 7
2.3 Average soil ingestion rate by age (From USEPA, 2011) .......... 8
2.4 Influences on soil Pb bioavailability (From Ruby et al., 1996) .... 10
2.5 Land categories in Part 2A of the EPA 1990 (From DEFRA, 2014) .... 13
2.6 Pb concentration by depth (From Walraven et al., 2014) ........ 17

3.1 Diagram of the G-BASE sampling strategy .......................... 29
3.2 Map of G-BASE sample locations .................................. 31
3.3 Map of Paisley study area with historic industry and G-BASE samples .... 34
3.4 Historic photo of Linwood Steel works, Paisley (From Canmore, n.d.) .... 33
3.5 Map of Bishopbriggs study area with build age & G-BASE samples ... 36
3.6 Historic photo of Springfield Road, Bishopbriggs (From Bishopbriggs History Group, 2021) ........................................ 35
3.7 Maps of sample locations in Paisley and Bishopbriggs ............... 38
3.8 Photographs showing sampling methodology .......................... 39
3.9 Concentration of M4 standard over time ............................. 45

4.1 Frequency histograms of soil Pb concentrations ....................... 55
4.2 Soil Pb vs coordinates in each study area ............................. 59
4.3 Multi-directional variograms for each study area ....................... 60

5.1 A semivariogram model (From Oliver and Webster, 2015) .......... 65
5.2 Correlograms for each study area ........................................ 78
5.3 Moran’s scatterplots for each study area ................................. 80
5.4 Map of samples coloured by Moran’s quadrant ......................... 81
5.5 REML semivariograms with MOM points for each study area ........ 83
5.6A Kriged soil Pb (Box-Cox) and kriging error for Greater Glasgow .... 85
5.6B Kriged soil Pb (Box-Cox) and kriging error for Paisley ............... 86
5.6C Kriged soil Pb (Box-Cox) and kriging error for Bishopbriggs ......... 87
5.7A Kriged soil Pb (mg kg⁻¹) for Greater Glasgow, error = transparency ... 88
5.7B Kriged soil Pb (mg kg⁻¹) for Paisley, error = transparency ............. 89
5.7C Kriged soil Pb (mg kg⁻¹) for Bishopbriggs, error = transparency ...... 90
5.8A Kriged soil Pb (classified by C4SL exceeded) for Greater Glasgow ... 91
5.8B Kriged soil Pb (classified by C4SL exceeded) for Paisley ............... 92
5.8C Kriged soil Pb (classified by C4SL exceeded) for Bishopbriggs ......... 93
5.9 Paisley kriged soil Pb using (A) G-BASE and (B) this study samples ... 99
5.10 Bishopbriggs kriged soil Pb using (A) G-BASE and (B) this study samples ......................................................... 100
6.1 Boxplots of soil Pb by land use in Dublin (From Glennon et al., 2014) 
6.2 Soil Pb with distance from a road, Dublin (From Dao et al., 2014) 
6.3 Lead levels in household dust and soil by property age in the UK (From Thornton et al., 1985) 
6.4 Soil, road, and household dust Pb levels for different aged houses, Brighton (Based on Davies and Thornton, 1986) 
6.5 Soil Pb levels around different aged houses in Baltimore, USA (From Schwarz et al., 2012) 
6.6 Soil Pb levels with building distance in Baltimore, USA (From Schwarz et al., 2012) 
6.7 Maps of (A) Pb domains and (B) interpolation soil Pb concentrations in England (From Ander et al., 2013) 
6.8 Maps of covariate distribution in Greater Glasgow 
6.9 Maps of covariate distribution in Paisley 
6.10 Maps of covariate distribution in Bishopbriggs 
6.11 Flow diagram summarising methodology for analysing covariates 
6.12 Box-plots and statistics of soil Pb by land use in Greater Glasgow 
6.13 Scatterplot of soil Pb vs road distance in Greater Glasgow 
6.14 Box-plots and statistics of soil Pb by traffic density in G. Glasgow 
6.15 Scatterplot of soil Pb vs historic industry distance in G. Glasgow 
6.16 Box-plots and statistics of soil Pb by historic ind. age in G. Glasgow 
6.17 Scatterplot of soil Pb vs historic industry density in G. Glasgow 
6.18 Box-plots and statistics of soil Pb by building age in G. Glasgow 
6.19 Scatterplot of soil Pb vs building distance in Greater Glasgow 
6.20 Scatterplot of soil Pb vs (A) OM and (B) pH in Greater Glasgow 
6.21 Box-plots and statistics of soil Pb by soil texture in G. Glasgow 
6.22 Scatterplot of soil Pb vs OM in (A) Paisley and (B) Bishopbriggs 
6.23 Box-plots of soil Pb by land use in (A) Paisley and (B) Bishopbriggs 
6.24 Scatterplot of soil Pb vs build dist. in (A) Paisley and (B) Bishopbriggs 
6.25 Scatterplot of soil Pb vs road dist. in (A) Paisley and (B) Bishopbriggs 
6.26 Box-plots of soil Pb by build age in (A) Paisley and (B) Bishopbriggs 
6.27 Box-plots of soil Pb by H.I. type in (A) Paisley and (B) Bishopbriggs 
6.28 Scatterplot of soil Pb vs H.I. type in (A) Paisley and (B) Bishopbriggs 
6.29 Map of Etive Crescent refuse heap, Bishopbriggs 
6.30 Map of historic railways and building age in Paisley 
6.31 Variograms modelling soil Pb in Paisley with different covariates 
6.32 Variograms modelling soil Pb in Bishopbriggs with covariates 
6.33 Maps of soil Pb (E-BLUP) with different covariates in Paisley 
6.34 Maps of soil Pb (E-BLUP) with different covariates in Bishopbriggs 
7.1 Three isotope plot of soil Pb (From Farmer et al., 2011)
7.2 Soil Pb from petrol and other sources with road distance, for different roads and years (From MacKinnon et al., 2011) ............ 191

7.3 Soil Pb from petrol and other sources away from roads along two transects in Glasgow (From Macintyre, 2017) ................... 192

7.4 Three isotope plot of soil Pb in Paisley and Bishopbriggs .......... 199

7.5 Kriged soil Pb % and conc. from petrol and coal sources in Paisley ........... 203

7.6 Kriged soil Pb % and conc. from petrol and coal in Bishopbriggs ....... 205

7.7 Scatterplots of road distance with (A) total and petrol Pb, (B) petrol Pb conc., (C) petrol Pb %, (D) coal conc., and (E) coal Pb % in Paisley Scatterplots of road distance with (A) total and petrol Pb, (B) petrol Pb conc., (C) petrol Pb %, (D) coal conc., and (E) coal Pb % in Bishopbriggs .......................................................... 207

7.8 Scatterplots of industry distance with (A) total and industry Pb, (B) petrol Pb conc., (C) petrol Pb %, (D) coal conc., and (E) coal Pb % in Paisley Scatterplots of industry distance with (A) total and industry Pb, (B) petrol Pb conc., (C) petrol Pb %, (D) coal conc., and (E) coal Pb % in Bishopbriggs .......................................................... 212

8.1 Example of a regression tree. (Based on Cutler et al., 2011) ............ 221

8.2 Flow diagram of method for training and validating a RF model ...... 233

8.3 Predictor importance plot of covariates in the G. Glasgow RF model . 235

8.4 G. Glasgow grid with median predicted soil Pb (log) from QRF ...... 236

8.5 G. Glasgow grid with (A) median predicted soil Pb, (B) the lower confidence interval, and (C) the upper confidence interval classified by C4SL exceeded at each point ....................................... 237

8.6 Map of the difference between the 95th percentile and median prediction in the Greater Glasgow prediction grid ................. 239

8.7 Predicted vs actual soil Pb (log) values in G. Glasgow for a range of percentiles and coloured by their soil Pb value (i.e. high and low) .................. 241

8.8 G. Glasgow grid with probability of having high (> 200 mg kg\(^{-1}\)) soil Pb from QRF predictions. With contour lines around hotspots .... 245

8.9 ROC curve, Youden’s index and AUC for G. Glasgow samples ....... 246

8.10 Confusion matrix for samples in Greater Glasgow ...................... 247

8.11 Maps of (A) observed and (B) predicted soil Pb group in G. Glasgow 248

8.12 PIP of covariates in (A) Paisley and (B) Bishopbriggs RF models .... 251

8.13 Median predicted Pb (log) in (A) Paisley and (B) Bishopbriggs .... 253

8.14 Paisley grid with (A) median predicted soil Pb, (B) the lower confidence interval, and (C) the upper confidence interval classified by C4SL exceeded at each point ................................. 254

8.15 Bishopbriggs grid with (A) median predicted soil Pb, (B) the lower confidence interval, and (C) the upper confidence interval classified by C4SL exceeded at each point ................................. 255
<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>8.16</td>
<td>Predicted vs actual soil Pb (log) values in Paisley for a range of percentiles and coloured by their soil Pb value (i.e. high and low)</td>
</tr>
<tr>
<td>8.17</td>
<td>Predicted vs actual soil Pb (log) values in Bishopbriggs for a range of percentiles and coloured by their soil Pb value (i.e. high and low)</td>
</tr>
<tr>
<td>8.18</td>
<td>(A) Paisley and (B) Bishopbriggs grids with probability of having high (&gt; 200 mg kg$^{-1}$) soil Pb from QRF predictions</td>
</tr>
<tr>
<td>8.19</td>
<td>ROC curves, YI, and AUC for Paisley and Bishopbriggs samples</td>
</tr>
<tr>
<td>8.20</td>
<td>Confusion matrices for samples in (A) Paisley and (B) Bishopbriggs</td>
</tr>
<tr>
<td>8.21</td>
<td>Maps of (A) observed and (B) predicted soil Pb group in Paisley</td>
</tr>
<tr>
<td>8.22</td>
<td>Maps of (A) observed &amp; (B) predicted soil Pb group in Bishopbriggs</td>
</tr>
<tr>
<td>9.1</td>
<td>Urban locations included in the G-BASE and Tellus projects</td>
</tr>
<tr>
<td>9.2A</td>
<td>Sample locations in Belfast</td>
</tr>
<tr>
<td>9.2B</td>
<td>Sample locations in Leicester</td>
</tr>
<tr>
<td>9.2C</td>
<td>Sample locations in Tamar</td>
</tr>
<tr>
<td>9.3</td>
<td>Predicted vs actual soil Pb (log) values in (A) Belfast and (B) Leicester</td>
</tr>
<tr>
<td>9.4</td>
<td>Maps of (A) observed and (B) predicted soil Pb group in Belfast</td>
</tr>
<tr>
<td>9.5</td>
<td>Maps of (A) observed and (B) predicted soil Pb group in Leicester</td>
</tr>
<tr>
<td>9.6</td>
<td>Maps of (A) observed and (B) predicted soil Pb group in Tamar</td>
</tr>
<tr>
<td>9.7</td>
<td>Leicester grid with probability of having high (&gt; 200 mg kg$^{-1}$) soil Pb from QRF predictions. With contour lines around hotspots</td>
</tr>
<tr>
<td>9.8</td>
<td>Confusion matrices for (A) Bishopbriggs predicted using the Paisley RF model and (B) Paisley using the Bishopbriggs RF model</td>
</tr>
<tr>
<td>9.9</td>
<td>PIP for RF models on (A) Glasgow, (B) Belfast, and (C) Leicester</td>
</tr>
<tr>
<td>9.10</td>
<td>PDP for (A) OM, (B) historic industry density, (C) historic industry type, and (D) land in the three urban datasets</td>
</tr>
<tr>
<td>9.11</td>
<td>Building age maps for each city and PDP for building age in each of the RF models</td>
</tr>
<tr>
<td>10.1</td>
<td>Summary of main map results from thesis</td>
</tr>
<tr>
<td>10.2</td>
<td>Summary of covariate importance</td>
</tr>
<tr>
<td>10.3</td>
<td>SIMD viewer focused on Glasgow (From SIMD.scot, 2020)</td>
</tr>
</tbody>
</table>
Tables

2.1 Soil Pb guideline values by country and land use

2.2 Examples of worldwide studies measuring urban soil Pb

3.1 Operating parameters used by ICP-OES

3.2 Concentrations of reagent blanks

3.3 Certified reference material concentrations

3.4 Diff. between measured and certified values for external checks

3.5 Duplicate and triplicate SD and RSD

3.6 Field duplicate results - difference in soil Pb concentrations

3.7 Comparison between G-BASE and my samples

3.8 Correlation coefficients of best-fit line from M4 standard solutions

3.9 Drift correction equations applied to samples for each ICP-OES run

4.1 Summary statistics of soil Pb concentrations in Glasgow study areas

4.2 Skew and transform statistics for each study area

5.1 Studies using variograms to measure soil Pb’s spatial variation

5.2 Summary of neighbours and distances in the spatial weights matrix

5.3 Observed and expected Global Moran’s I values for each study area

5.4 Summary of LISA statistics for each study area

5.5 Semivariogram model parameters for each study area

5.6 Cross validation statistics for REML fit variogram models

5.7 Comparison of semivariogram fitting methods and platforms

6.1 Summary of main covariates influencing soil Pb in urban areas

6.2 Studies comparing soil Pb concentrations by land use

6.3 Studies measuring soil Pb values by road distance and traffic density

6.4 Industrial processes Pb emissions (From Pacyna and Pacyna, 2000)

6.5 A summary of industrial processes producing Pb

6.6 Summary of results from covariate analysis in Greater Glasgow. (A) summary statistics, (B) model validation statistics, and (C) the relationship with soil Pb

6.7 Historical industry aggregation and soil Pb levels for each group

6.8 Building age group justifications and maps used to digitise

6.9 Individual Wald test results for building age groups

6.10 Summary of Wald and p values for each covariate in G. Glasgow

6.11 Summary of results from covariate analysis in Paisley and Bishopbriggs. (A) summary statistics, (B) variogram parameters, (C) model validation statistics, and (D) the relationship with soil Pb

6.12 Summary of Wald and p values for covariates in each study area
Equations

3.1 Lead concentration (mg kg⁻¹) .................................................. 42
3.2 Organic matter (OM) percentage ............................................... 43
3.3 Limit of detection (LOD) .......................................................... 43
3.4 Recovery % of CRM ............................................................... 44
3.5 Difference between duplicates .................................................. 46
3.6 Mean absolute deviation ......................................................... 46

4.1 Octile skew ................................................................. 53
4.2 Box-Cox transform ............................................................. 54

5.1 Semivariance ............................................................... 64
5.2 Observed Moran’s I ............................................................. 70
5.3 Expected Moran’s I ............................................................. 71
5.4 Observed Local Moran’s I (LISA) ............................................. 71
5.5 Expected Local Moran’s I (LISA) ............................................. 71
5.6 Linear mixed model (LMM) .................................................... 73
5.7 Likelihood function ............................................................... 73
5.8 Restricted maximum likelihood (REML) ................................... 73
5.9 Ordinary kriging (OK) .......................................................... 74
5.10 Standardised squared error (SSE) ........................................... 75

6.1 Akaike Information Criteria (AIC) ........................................... 139
6.2 Wald test null hypothesis ....................................................... 140
6.3 Wald test contrast matrix ....................................................... 140
6.4 Wald statistic ................................................................. 140
6.5 The empirical best linear unbiased predictor (E-BLUP) ......... 141

7.1 Soil Pb % from petrol and coal sources ................................... 194

8.1 Random forest training ....................................................... 220
8.2 Random forest prediction .................................................... 222
8.3 Root mean square error (RMSE) ........................................... 231
8.4 Concordance correlation coefficient (CCC) ......................... 231
8.5 Youden’s index (YI) ............................................................ 232
Chapter 1. Introduction

1.1. Project background

This project originated as PRELUDE (PREdictive modelling of Lead concentrations using G-BASE datasets for Urban Environments) and aimed to develop a model that can estimate soil lead (Pb) concentrations in Greater Glasgow using the British Geological Survey’s (BGS) Geochemical Baseline Survey of the Environment (G-BASE) dataset(s). Schwarz et al.’s (2013) paper sparked this idea, where predictor variables (i.e. road distance, housing age, and housing distance) were incorporated into three predictive models (i.e. a general linear model (GLM), classification and regression tree (CART), and random forest (RF) model). These models determined if soil in Baltimore, USA, would be above or below a 400 mg kg\textsuperscript{-1} threshold for Pb concentration and showed a reasonable degree of accuracy (>72%).

Predictive modelling of soil Pb concentrations can save time and money by focusing sampling investigation, and any resulting future remediation, to the most at-risk areas. This is based on the premise that predictive models generate a map of soil Pb concentrations throughout a city by building on existing datasets e.g. the G-BASE survey. Specifically, the relationship between soil Pb concentrations and digital datasets of predictor variables (e.g. road distance, housing age, historic industry distance) can be modelled, and used to determine soil Pb concentrations in unsampled locations. The locations posing the greatest potential health threat can then be investigated, and even remediated, at a lower overall cost.

The Schwarz et al. (2013) study demonstrated this potential benefit of predictive models, but also indicated points for development. For instance, similar models predicting soil Pb concentration have not been tested on a British city, where the developmental history and individual factors can vary from American cities. Additionally, the Schwarz et al. models were not extensively validated, as the independent validation dataset was limited, and the models were not tested on other American cities. Furthermore, model accuracy was not examined at different scales, for example it is unknown whether the model trained on data at the city scale will give accurate predictions at the neighbourhood scale. There was also no measure of uncertainty in their models. Lastly, only three predictor variables were input in Schwarz et al.’s models, and it is likely that other factors, such as historic industry location, will influence the distribution of soil Pb concentrations.
1.2. Research questions

Given the above points, this thesis aims to answer six research questions:

1. How important is scale when mapping and modelling soil Pb concentrations?
2. Can the sources of Pb in an urban environment be ascertained by exploring the relationships between sampled soil Pb chemistry and landscape features?
3. Are models (both geostatistical and machine learning) a viable method for accurately predicting soil Pb concentrations?
4. Are machine learning models developed for one city applicable to other cities and geographical locations?
5. Can these machine learning models improve our understanding of areas that present a greater risk from soil Pb?
6. Which methods are most suitable for communicating the results of such models?

1.3. Thesis outline

To address these questions the thesis is split into ten chapters. The next chapter will examine the current literature on urban soil Pb, focusing on the health implications from Pb, how soil ingestion/inhalation acts as an exposure pathway, some recommended thresholds of Pb in soils, the importance of studying urban environments, and why soil Pb concentrations are still an important issue. These topics will justify the necessity of studying soil Pb in urban environments.

The third chapter will outline the methodology used to obtain datasets on soil Pb concentrations. As scale is an important topic throughout this thesis, soil Pb concentration datasets are required at the city scale and at the neighbourhood scale. The BGS’s G-BASE dataset covering Greater Glasgow will be used for city-wide measurements of total soil Pb concentrations, along with organic matter (OM) content and soil pH. More spatially detailed samples, collected from two neighbourhoods in Glasgow (Paisley and Bishopbriggs), and analysed in the laboratory, will provide total soil Pb concentrations, OM, and pH data at the finer scale.

In Chapter 4 exploratory data analysis will be performed in the three study areas. This will determine average soil Pb concentrations and assess if their distribution is
skewed or shows an isotropic pattern. A skewed distribution or isotopic pattern will require a data transform before further analysis.

Chapter 5 explores differences in the spatial dependence and variability of soil Pb concentrations between the city and neighbourhood scales. This will address the first research question i.e. how important is scale when modelling and mapping soil Pb concentrations? Moran’s I analysis will assess spatial dependence, and semivariograms will measure the spatial range over which such dependence occurs. Lastly, ordinary kriging will map soil Pb concentrations, at both the city and neighbourhood scale.

Spatial patterns in soil Pb concentrations, revealed by the kriged maps, will then be investigated. Specifically, Chapter 6 seeks to understand the sources of soil Pb at the city and neighbourhood scale, which will answer research question two. This topic is also explored in Chapter 7, where Pb isotopes will be used to apportion Pb originating from petrol and industrial (or coal) sources. Determining the most important sources of soil Pb will be vital before predictive modelling is undertaken. This is because only covariates with a recognised relationship with soil Pb concentrations should be input in future predictive models, to avoid the issues of data duplication and model over-fitting.

The third research question – are machine learning models a viable method for accurately predicting soil Pb concentrations – will be tackled in Chapter 8. In this chapter, a machine learning technique called random forests will be used to predict soil Pb concentrations at the city and neighbourhood scale. Model accuracy will be determined by cross-validation. Chapter 8 will also examine how to map soil Pb risk, and the most suitable method for communicating the results (research questions five and six).

In Chapter 9 the models generated will be tested on other UK cities and regions, where an independent validation dataset is present (thereby addressing research question four). This is the first time a random forest model will be evaluated at predicting soil Pb concentrations in a city it was not trained for.

Lastly, Chapter 10 will present an overall summary of the key findings from each chapter, and a more in-depth discussion of the results taken together, along with considering their wider significance. The final chapter will also discuss further research opportunities arising from this project.
All the work carried out in this thesis was undertaken by myself. This includes the initial analysis of the G-BASE dataset, digitisation of historic maps, soil sampling in Paisley and Bishopbriggs, laboratory analysis (for total Pb concentrations and isotopic Pb) in all 250 neighbourhood samples, analysis of the neighbourhood datasets, and training and testing of RF models. However, the BGS kindly provided the G-BASE dataset, and Dr Benjamin Marchant supplied initial Matlab code which I adapted.
Chapter 2: Literature review on urban soil Pb

Understanding the potential dangers posed by Pb in urban soils is essential for mitigating against these risks. Otherwise, there is little point in mapping soil Pb concentrations and risk using predictive modelling. It is also important to appreciate recommended maximum safe concentrations for Pb in urban soils, in order to determine soil Pb risk in later chapters.

2.1. Health implications associated with Pb

Lead is a naturally occurring metal, which is toxic to most organisms (Steinnes, 2012). However, Pb can enter the human bloodstream through multiple exposure routes (Section 2.2). Elevated human blood Pb levels are concerning because Pb has been associated with adverse health effects. These adverse effects have been observed for millennia (Needleman, 2009). For example, Egyptian scrolls describe the usefulness of Pb compounds for homicide (Hernberg, 2000), and the Romans love of Pb for coating pottery and sweetening wine likely contributed to their low fertility, psychosis, and perhaps even the fall of Rome (Needleman, 2004). More recent research has identified that children, especially those under six years old, are more susceptible to Pb toxicity (e.g. King, 1971; Canfield et al., 2003; Mielke et al., 2010; Naranjo et al., 2020). This is due to children’s developing nervous system (Rice and Barone, 2000), extended contact with floors and soil, hand-mouth and pica tendencies (Needleman, 2009; Elom et al., 2013), and higher gastrointestinal absorption of Pb into the bloodstream (Appleton et al., 2012).

The neurological effects of Pb are well-documented and even low Pb exposure levels in humans can cause central nervous system and neurological damage (Lanphear et al., 2005). The effect is most commonly a drop in intelligence quotient (IQ; Figure 2.1; Canfield et al., 2003; Jakubowski, 2011; Skerfving et al., 2015; Galiciolli et al., 2022), and was the reason the American Centres for Disease Control and Prevention (CDC) reduced the recommended limit of blood Pb from 50 µg dL\(^{-1}\) to 3.5 µg dL\(^{-1}\) over the past 60 years (CDC, 2022). However, blood Pb concentrations below this threshold may still pose a risk to health (Canfield et al., 2003). Overall, there has been an observed decrease in human blood Pb levels over the past 30 years. This is due to the removal of Pb from some sources humans may be exposed to e.g. petrol, paint, and water pipes. However, blood Pb levels
have not yet decreased to a ‘safe’ level where no health effects are observed (Etchevers et al., 2014; Naranjo et al., 2020).

Lead-induced neurological damage in young children can also influence their behaviour later in life (Zeng et al., 2021). Specifically, exposure to Pb at a young age has been linked with impulsive and aggressive behaviour (Reyes, 2015), a reduction in language abilities (Koshy et al., 2020), criminal tendencies (Wright et al., 2008; Mielke and Zahran, 2012), and attention deficit and hyperactivity disorders (Nicolescu et al., 2010; Kim et al., 2013). Lead does not just affect the brain, but accumulates in the blood, soft tissues and eventually the bones (Cornelis et al., 2005). This can cause hearing impairment (Osman et al., 1999) and osteoporosis (Campbell et al., 2004) in children, and in adults may contribute to hypertension (Cheng et al., 2001), premature birth in pregnant woman (Khanam et al., 2021), stroke mortality (Menke et al., 2006), reduced male fertility (Hernández-Ochoa et al., 2005), anaemia (Breeher et al., 2013), cataracts (Schaumberg et al., 2004), and renal diseases (Tsaih et al., 2004).

**Figure 2.1:** IQ versus blood Pb for all age groups (From Canfield et al., 2003). Dashed line indicates the current recommend limit for blood Pb (3.5 µg dl⁻¹: CDC, 2022).
2.2. Soil ingestion/inhalation as an exposure pathway

For these adverse health effects to occur, Pb must first enter the human bloodstream. There are multiple potential exposure pathways of humans to Pb, which includes ingestion of food and/or water, inhalation of air and dermal adsorption of Pb (EA, 2011). Migration of soil Pb into the body is an important pathway and could arise from several routes (Figure 2.2). These pathways are often controlled by soil Pb bioavailability i.e. the amount of Pb freely available to cross an organism’s cellular membrane from the soil, and soil Pb bioaccessibility i.e. the total amount of Pb that is available to an organism/plant for uptake, but the Pb may be physically constrained (Semple et al., 2004).

Figure 2.2: Main exposure pathways of Pb to children via soil. Arrow thickness represents relative movement of Pb between pools. Figure made using references in Section 2.2.
Ingestion is a major exposure pathway to soil Pb, especially for children (Figure 2.2; Lanphear and Roghman, 1997; Laidlaw and Filippelli, 2008; Qu et al., 2012). Soil may be ingested intentionally by young children who display soil-pica i.e. ingest more than 1,000-5,000 mg of soil per day (U.S. Environmental Protection Agency (USEPA), 2011; Elom et al., 2013). However, unintentional ingestion of Pb contaminated soil is a more common exposure pathway. This occurs because Pb associated with fine soil particles (<63 μm) readily adheres to children’s hands (Choate et al., 2006; Wang et al., 2023) and is likely to enter the gastrointestinal system due to children’s mouthing tendencies (Moya and Phillips, 2014). Lead associated with finer particles also tends to be more bioaccessible (Wang et al., 2023). This is more prevalent in toddlers, as crawling puts them into direct contact with soil or floors (Zahran et al., 2013). Subsequently, the average amount of soil ingested per day varies with age (Figure 2.3; USEPA, 2011).

![Estimated Daily Soil Ingestion](image)

**Figure 2.3:** USEPA daily soil ingestion estimates for different age groups. Soil pica is not accounted for, and the overall confidence rating of these recommendations are low due to significant methodology limitations (From USEPA, 2011).

Another possible soil exposure pathway is via vegetables or crops grown in Pb contaminated soil. This is especially relevant due to the growing popularity of urban agriculture, i.e. growing food in gardens, on vacant ground, or unusual city spaces (Redwood, 2012; Taylor and Lovell, 2012; Lal, 2020). However, the importance of this exposure pathway is highly controversial. Typically, Pb has low bioavailability, i.e. low concentrations of free Pb²⁺ ions in the soil are transferred to plant roots (Shahid et al., 2012). Furthermore, once Pb reaches plant roots only very small amounts are translocated to the edible parts of the plant due to physical
obstruction by the Casparian strip (Gao et al., 2011; Pourrut et al., 2011). Therefore, only root vegetables (e.g. carrots and potatoes) have been shown to contain Pb levels above the maximum guideline value for Pb, whereas leafy vegetables and crops tend to contain safe Pb levels (Douay et al., 2013; Attanayake et al., 2014). A review by Brown et al. (2015) concluded that the risks of elevated soil Pb concentrations to urban agriculture is negligible. Aerial components of vegetation accumulate more atmospheric Pb via stomata openings and cuticle diffusion, known as foliar uptake, compared with root uptake from soil sources, but this is still minimal (Uzu et al., 2011; Schreck et al., 2012; Bidar et al., 2020). A more likely pathway is ingestion of soil or dust that are not washed off fruits or vegetables before consumption (Douray et al., 2013).

Lead inhalation is another potential exposure route and can occur via direct inhalation of Pb associated with particulate matter (e.g. PM$_{10}$) or as a consequence of resuspended legacy Pb from the soil, which is often tracked indoors via air currents, shoes, or pets (Lanphear and Roghmann, 1997; Laidlaw and Filippelli, 2008). Contemporary Pb inhalation is predominantly due to resuspension of legacy Pb, whereas in the past Pb was inhaled directly from petrol emissions or industrial sources (Zahran et al., 2013). After inhalation, some of the Pb associated with particulate matter is absorbed to the bloodstream through the tracheobronchial tract, specifically the pulmonary alveoli (Julien et al., 2011). Generally, Pb has low respiratory bioaccessibility; however, this is highly variable due to differences in Pb speciation, and particle size and/or shape, which are predominantly controlled by the Pb source (Voutsa and Samara, 2002; Julien et al., 2011; Boisa et al., 2014). For instance, Pb associated with finer particulates (< 2.1 μm) tends to originate from petrol emissions and has a greater lung penetration capacity compared with coarser particles (2.1 – 10 μm) from industrial sources (Julien et al., 2011; Avino et al., 2013; Jia et al., 2022).

Similarly, not all ingested or inhaled soil Pb enters the bloodstream i.e. is bioaccessible. Studies have shown that on average only ~50% of Pb is bioaccessible at the gastric phases, which drops to 25% bioaccessibility from the gastrointestinal tract (e.g. Farmer et al., 2011; Luo et al., 2012a; Pelfrène et al., 2013; Reis et al., 2013; Kelepertzis et al., 2015). However, bioaccessibility is often influenced by land use (Pelfrène et al., 2013), the original Pb source (Walraven et al., 2015), soil properties (Cave et al., 2012), and the predominant soil phases.
associated with Pb (Farmer et al., 2011; Cai et al., 2015). The longer Pb resides in the soil the less impact source has on bioaccessibility, and Pb speciation and associations become more significant (Walraven et al., 2015). Specifically, Pb associated with sulphides, clays, OM, and reactive iron (Fe) have low Pb bioaccessibility as they form more stable complexes which are harder to absorb (Figure 2.4; Madrid et al., 2008; Farmer et al., 2011; Cai et al., 2015; Walraven et al., 2015). Conversely, Pb bioaccessibility is highest for Pb associated with carbonates and oxides, as it is more exchangeable in these forms (Figure 2.4; Ruby et al., 1996; Farmer et al., 2011). The relationship between Pb and its associated soil components has been examined in numerous studies using multi step sequential extraction (Gibson and Farmer, 1986; Bacon et al., 2006; Farmer et al., 2011; Rinklebe and Shaheen, 2014).

**Figure 2.4:** Depiction of how the mineral form, particle size, and encapsulation impact Pb bioavailability (From Ruby et al., 1996).
2.3. How much soil Pb is too much?

The health implications and exposure pathways associated with soil Pb have prompted many countries to determine the maximum ‘safe’ level of Pb in the soil i.e. a soil guideline value (SGV). These SGVs are often based on a ‘safe’ blood Pb level, so have been lowered over time due to further understanding of Pb migration and its effects (Table 2.1). The SGVs also vary between countries, are dependent on land use and other soil factors, and are highly controversial due to their legal status (Table 2.1).

<table>
<thead>
<tr>
<th>Country</th>
<th>Publication</th>
<th>Date</th>
<th>Land Use</th>
<th>Pb SGV (mg kg⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>UK</strong></td>
<td>ICRL 59/83 trigger concentrations</td>
<td>1987-2002</td>
<td>Domestic gardens, allotments</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Parks, playing fields, open space</td>
<td>2000</td>
</tr>
<tr>
<td></td>
<td>DEFRA SGVs (from CLEA)</td>
<td>2002</td>
<td>Domestic gardens, allotments</td>
<td>450</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commercial, industrial</td>
<td>750</td>
</tr>
<tr>
<td></td>
<td>DEFRA Category 4 Screening Levels (C4SLs)</td>
<td>2014</td>
<td>Allotments</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Residential (with home-grown produce)</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Residential (no home-grown produce)</td>
<td>310</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Public Open Space (residential)</td>
<td>630</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Public Open Space (park)</td>
<td>1300</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commercial</td>
<td>2300</td>
</tr>
<tr>
<td><strong>Netherlands</strong></td>
<td>Dutch Target and Intervention values</td>
<td>2000</td>
<td>Standard soil (10% OM, 25% clay) Intervention value</td>
<td>530</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Target value</td>
<td>85</td>
</tr>
<tr>
<td><strong>USA</strong></td>
<td>US Environmental Protection Agency soil screening value (SSV)</td>
<td>2005</td>
<td>Residential - where children play</td>
<td>400</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Residential - rest of yard</td>
<td>1200</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commercial, industrial</td>
<td>3500</td>
</tr>
<tr>
<td></td>
<td>California revised SSV</td>
<td>2009</td>
<td>Residential</td>
<td>80</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commercial, industrial</td>
<td>320</td>
</tr>
<tr>
<td><strong>New Zealand</strong></td>
<td>Ministry for the Environment - soil standards</td>
<td>2011</td>
<td>Residential (25% home-grown produce)</td>
<td>160</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Residential (no home-grown produce)</td>
<td>250</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>High density residential</td>
<td>500</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Recreational</td>
<td>880</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Commercial, industrial</td>
<td>3300</td>
</tr>
</tbody>
</table>
In the UK, the Inter-departmental Committee for the Redevelopment of Contaminated Land (ICRCL) was founded in 1976 to investigate land contamination (Nathanail, 2011). The ICRCL published trigger values for Pb (ICRCL 59/83) in 1987, which were based on site investigations, and were dependent on the future use of the site (Table 2.1).

In 2002 the ICRCL’s trigger values were replaced with lower SGVs from the Department for Environment, Food and Rural Affairs (DEFRA; Cole and Jeffries, 2009). However, unlike the ICRCL’s trigger values these SGVs were not statutory, instead they supported Part 2A of the Environmental Protection Act (Cole and Jeffries, 2009). Part 2A of the Environmental Protection Act of 1990 was formally introduced in Scotland in 2000 and provides a legal framework for the definition and remediation of contaminated land (DEFRA, 2012). This legislation is risk based, in that contaminated land is defined as land which:

"... appears to the local authority... to be in such a condition by reason of substances in, on, or under the land that (a) significant harm [to human health] is being caused or there is a significant possibility of such harm being caused...” (DEFRA, 2012).

DEFRA’s definition of ‘significant harm’ is dependent on source-pathway-receptor links. As discussed previously (Section 2.2.), these exposure pathways are often site dependent. Therefore, the Environmental Agency’s Contaminated Land Exposure Assessment (CLEA) conceptual modelling tool is used to determine if an area of land poses a significant possibility of harm (Jeffries, 2009). CLEA combines information on land use and soil properties to compare the predicted exposure with health criteria values e.g. blood Pb levels (Jeffries, 2009).

General assumptions were applied to the CLEA model to generate the SGVs for Pb (Table 2.1). Therefore, the SGVs are considered a simplified assessment of human health risk (Cole and Jeffries, 2009). Crucially, not all instances of Pb concentrations above the SGV will impact human health, there is only the potential to cause harm (Davies, 2008). Consequently, sites with Pb values exceeding the SGV need to be investigated in more detail i.e. run a source-pathway-receptor risk assessment to determine whether interventions are required (Fordyce et al., 2012).

Subsequently, DEFRA has adopted Category 4 Screening Levels (C4SLs) which supersede the SGVs. A Steering group modified toxicological and exposure
pathways in the CLEA model to classify land as ‘low risk’ or category 4 if it is safe to use (DEFRA, 2014). Any land belonging in categories 1 or 2 are legally classified as contaminated, under the new guidance for Part 2A of the Environmental Protection Act (Figure 2.5; DEFRA, 2014). These C4SLs have not yet been adopted in Scotland, and the Scottish Government is still undertaking a risk assessment project, where they aim to determine levels of Pb, and other potentially harmful elements, which could cause the significant possibility of significant harm (CL:AIRE, 2022).

![Figure 2.5: Land categories described in the revised guidance for Part 2A of the Environmental Protection Act 1990. SGVs and C4SL are added. From DEFRA, 2014.](image)

2.4. The importance of studying Pb in urban environments

Lead is a naturally occurring element with an average concentration of 15 mg kg\(^{-1}\) in the Earth’s crust (Heinrichs et al., 1980; Wedepohl, 1995). However, geogenic Pb is unevenly distributed. In igneous rocks Pb concentrations are related to silica (Si) content, with felsic rocks e.g. granites (mean Pb = 23 mg kg\(^{-1}\)), exhibiting higher Pb concentrations than mafic rocks e.g. gabbros (mean Pb = 2 mg kg\(^{-1}\); Hooda, 2010). In sedimentary rocks Pb is associated with OM and sulphides, so tends to be highest in black shales (mean Pb = 30 mg kg\(^{-1}\)) compared with sandstones (mean Pb = 12 mg kg\(^{-1}\)) and limestones (mean Pb = 9 mg kg\(^{-1}\); Heinrichs et al., 1980; Hooda, 2010). Pb is a chalcophile element; hence, is primarily found in lead-sulphide (Pb-S) ores such as galena, which contains high levels of Pb (~83%; Thornton et al., 2001; Palero-Fernandez and Martin-Izard, 2005). Primarily, the soil
Pb concentration will reflect the composition of the underlying bedrock, so will vary considerably. In addition, the presence of secondary Fe and manganese (Mn) oxides can trap Pb in soil and Pb is often elevated in peat-type soil where Pb deposited from the atmosphere is bound by OM (e.g. Macgregor et al., 2015). However, there is a stronger link between bedrock Pb and soil Pb in rural environments, as urban soils tend to be highly disturbed by anthropogenic activity (Thornton et al., 2001).

Owing to the large spatial variation in soil Pb, studies have attempted to determine the mean or background concentration of Pb in natural topsoils. In England and Wales background Pb levels have been quoted at 41 mg kg\(^{-1}\) (Ander et al., 2013), whereas values are consistently lower if only Scottish soils are considered i.e. mean Pb = 32 mg kg\(^{-1}\) for mineral soils (Paterson et al., 2011). Furthermore, studies examining worldwide soil Pb concentrations concluded that background Pb levels are around 25 mg kg\(^{-1}\) (Matschullat et al., 2000).

Soils in urban areas worldwide often exhibit Pb concentrations well above this natural background concentration of 25 mg kg\(^{-1}\) (Table 2.2). This is due to undisputable evidence linking elevated Pb in urban environments with anthropogenic sources (Kelly et al., 1996; Farmer et al., 2011; Li et al., 2011; Luo et al., 2015, McIlwaine et al., 2017).

By 2050, two thirds of the world’s population are projected to inhabit urban areas, in 1950 this value was just one third (UN, 2014). This substantial growth in urban populations, especially in more developed nations, has caused increasing concern about the state of urban soils. Consequently, the majority of research into urban soils has occurred over the past 30 years and has focused on the health impacts associated with soil pollution (e.g. early studies include Mielke 1991; Thornton, 1991; Tiller, 1992). Urban soils form via complex processes, distinct from natural paedogenesis (Ajmone-Marsan and Biasioli 2010). Therefore, they are highly heterogeneous, both temporally because of rapid land use change, and spatially due to soil sealing and human disturbance (Ajmone-Marsan and Zanini, 2013). This makes urban soils more challenging to study than their rural counterparts.
Table 2.2: Examples of studies measuring Pb soil concentrations in urban areas worldwide, sorted by continent then alphabetically by country. All numbers are rounded to the closest whole number and soil Pb data have been converted to mg kg\(^{-1}\). Populations are rounded to the nearest 1,000 and are taken from the studies where possible and from UN data where not (UN, 2014). ICP-MS/A/OES = inductively coupled plasma-mass spectrometry/atomic/optical emission spectroscopy; (P)XRF/D = (portable) X-ray fluoresce/diffraction; (G)FAAS = (graphite) furnace atomic absorption spectroscopy.

<table>
<thead>
<tr>
<th>City, Country</th>
<th>Population</th>
<th>Land use</th>
<th>Number of samples</th>
<th>Analysis method</th>
<th>Pb (mg kg(^{-1}))</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>World background</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>-</td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Jakobstad, Finland</td>
<td>20,000</td>
<td>Urban</td>
<td>32</td>
<td>ICP-MS</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Athens, Greece</td>
<td>664,000</td>
<td>Urban</td>
<td>45</td>
<td>ICP-MS</td>
<td>157</td>
<td>159</td>
</tr>
<tr>
<td>Galway, Ireland</td>
<td>80,000</td>
<td>Urban</td>
<td>166</td>
<td>ICP-AES</td>
<td>78</td>
<td>72</td>
</tr>
<tr>
<td>Palermo, Italy</td>
<td>656,000</td>
<td>Parks</td>
<td>70</td>
<td>XRF</td>
<td>-</td>
<td>111</td>
</tr>
<tr>
<td>Salerno, Italy</td>
<td>138,000</td>
<td>Urban</td>
<td>151</td>
<td>ICP-AES</td>
<td>126</td>
<td>172</td>
</tr>
<tr>
<td>Torino, Italy</td>
<td>900,000</td>
<td>Urban</td>
<td>20</td>
<td>FAAS</td>
<td>277</td>
<td>397</td>
</tr>
<tr>
<td>Bergen, Norway</td>
<td>240,000</td>
<td>Nurseries</td>
<td>435</td>
<td>ICP-AES</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Oslo, Norway</td>
<td>550,000</td>
<td>Urban</td>
<td>300</td>
<td>ICP-AES</td>
<td>56</td>
<td>-</td>
</tr>
<tr>
<td>Trondheim, Norway</td>
<td>164,000</td>
<td>Urban</td>
<td>321</td>
<td>ICP-OES</td>
<td>81</td>
<td>-</td>
</tr>
<tr>
<td>Lodz City, Poland</td>
<td>750,000</td>
<td>Urban</td>
<td>238</td>
<td>?</td>
<td>56</td>
<td>80</td>
</tr>
<tr>
<td>Aveiro, Portugal</td>
<td>70,000</td>
<td>Urban</td>
<td>20</td>
<td>FAAS</td>
<td>72</td>
<td>46</td>
</tr>
<tr>
<td>Lisbon, Portugal</td>
<td>531,000</td>
<td>Gardens</td>
<td>51</td>
<td>ICP-MS</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ljubljana, Slovenia</td>
<td>270,000</td>
<td>Urban</td>
<td>20</td>
<td>FAAS</td>
<td>102</td>
<td>56</td>
</tr>
<tr>
<td>Seville, Spain</td>
<td>700,000</td>
<td>Urban</td>
<td>20</td>
<td>FAAS</td>
<td>223</td>
<td>284</td>
</tr>
<tr>
<td>Aberdeen, UK</td>
<td>202,000</td>
<td>Urban</td>
<td>14</td>
<td>GFAAS</td>
<td>117</td>
<td>-</td>
</tr>
<tr>
<td>Cardiff, UK</td>
<td>325,000</td>
<td>Urban</td>
<td>508</td>
<td>XRF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Glasgow, UK</td>
<td>592,000</td>
<td>Urban</td>
<td>1285</td>
<td>XRF</td>
<td>198</td>
<td>357</td>
</tr>
<tr>
<td>London, UK</td>
<td>8,539,000</td>
<td>Urban</td>
<td>6245</td>
<td>XRF</td>
<td>297</td>
<td>381</td>
</tr>
<tr>
<td>Manchester, UK</td>
<td>2,550,000</td>
<td>Urban</td>
<td>300</td>
<td>XRF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Scunthorpe, UK</td>
<td>730,000</td>
<td>Urban</td>
<td>196</td>
<td>XRF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Sheffield, UK</td>
<td>552,000</td>
<td>Urban</td>
<td>569</td>
<td>XRF</td>
<td>244</td>
<td>-</td>
</tr>
<tr>
<td>Swansea, UK</td>
<td>241,000</td>
<td>Urban</td>
<td>373</td>
<td>XRF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wolverhampton UK</td>
<td>250,000</td>
<td>Urban</td>
<td>285</td>
<td>XRF</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>City, Country</td>
<td>Population</td>
<td>Land use</td>
<td>Number of samples</td>
<td>Analysis method</td>
<td>Pb (mg kg⁻¹)</td>
<td>Reference</td>
</tr>
<tr>
<td>--------------</td>
<td>------------</td>
<td>-----------</td>
<td>-------------------</td>
<td>-----------------</td>
<td>---------------</td>
<td>---------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>The Americas</td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ottawa, Canada</td>
<td>800,000</td>
<td>Urban</td>
<td>50</td>
<td>ICP-MS</td>
<td>65</td>
<td>34</td>
</tr>
<tr>
<td>Talcahuano, Chile</td>
<td>163,000</td>
<td>Schools</td>
<td>76</td>
<td>ICP-ES</td>
<td>35.2</td>
<td>43.3</td>
</tr>
<tr>
<td>Baltimore, MD, USA</td>
<td>622,000</td>
<td>Urban</td>
<td>1,121</td>
<td>PXRF</td>
<td>363</td>
<td>7</td>
</tr>
<tr>
<td>Aspen, CO, USA</td>
<td>7,000</td>
<td>Gardens</td>
<td>65</td>
<td>ICPS</td>
<td>172</td>
<td>155</td>
</tr>
<tr>
<td>New Orleans, LA, USA</td>
<td>379,000</td>
<td>Urban</td>
<td>128</td>
<td>?</td>
<td>272</td>
<td>108</td>
</tr>
<tr>
<td>Miami, FL, USA</td>
<td>418,000</td>
<td>Urban</td>
<td>204</td>
<td>GFAAS</td>
<td>92</td>
<td>3</td>
</tr>
<tr>
<td>Syracuse, NY, USA</td>
<td>145,000</td>
<td>Urban</td>
<td>194</td>
<td>ICPS</td>
<td>80</td>
<td>-</td>
</tr>
<tr>
<td>Greensboro, NC, USA</td>
<td>300,000</td>
<td>Gardens</td>
<td>2310</td>
<td>PXRF</td>
<td>121</td>
<td>110</td>
</tr>
<tr>
<td>Africa</td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Annaba, Algeria</td>
<td>621,000</td>
<td>Urban</td>
<td>101</td>
<td>FAAS</td>
<td>53</td>
<td>85</td>
</tr>
<tr>
<td>Addis Ababa, Ethiopia</td>
<td>3,385,000</td>
<td>Urban</td>
<td>7</td>
<td>ICP-MS</td>
<td>11</td>
<td>3</td>
</tr>
<tr>
<td>Abuja, Nigeria</td>
<td>776,000</td>
<td>Urban</td>
<td>40</td>
<td>FAAS</td>
<td>281</td>
<td>39</td>
</tr>
<tr>
<td>Osofo, Nigeria</td>
<td>280,000</td>
<td>Roadside</td>
<td>39</td>
<td>FAAS</td>
<td>38</td>
<td>17</td>
</tr>
<tr>
<td>Kampala, Uganda</td>
<td>1,600,000</td>
<td>Roadside</td>
<td>52</td>
<td>FAAS</td>
<td>45</td>
<td>6</td>
</tr>
<tr>
<td>Asia</td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Changxing, China</td>
<td>620,000</td>
<td>Mixed</td>
<td>665</td>
<td>ICP-MS</td>
<td>33</td>
<td>7</td>
</tr>
<tr>
<td>Changchun, China</td>
<td>3,580,000</td>
<td>Urban</td>
<td>352</td>
<td>XRF</td>
<td>35</td>
<td>26</td>
</tr>
<tr>
<td>Guangzhou, China</td>
<td>13,080,500</td>
<td>Urban</td>
<td>426</td>
<td>AAS</td>
<td>88</td>
<td>238</td>
</tr>
<tr>
<td>Hong Kong, China</td>
<td>7,188,000</td>
<td>Urban</td>
<td>40</td>
<td>ICP-AES</td>
<td>108</td>
<td>89</td>
</tr>
<tr>
<td>Rizhao, China</td>
<td>1,360,000</td>
<td>Urban</td>
<td>445</td>
<td>FAAS</td>
<td>28</td>
<td>6</td>
</tr>
<tr>
<td>Xiamen, China</td>
<td>3,730,000</td>
<td>Parks</td>
<td>39</td>
<td>ICP-OES</td>
<td>32</td>
<td>-</td>
</tr>
<tr>
<td>Mumbi, India</td>
<td>12,436,000</td>
<td>Urban</td>
<td>?</td>
<td>XRD</td>
<td>42</td>
<td>62</td>
</tr>
<tr>
<td>Bangkok, Thailand</td>
<td>5,888,000</td>
<td>Urban</td>
<td>30</td>
<td>GFAAS</td>
<td>48</td>
<td>53</td>
</tr>
<tr>
<td>Australasia</td>
<td></td>
<td></td>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Median</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Range</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Reference</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adelaide, Australia</td>
<td>1,076,000</td>
<td>Urban</td>
<td>69</td>
<td>?</td>
<td>97</td>
<td>49</td>
</tr>
<tr>
<td>Adelaide, Australia</td>
<td>1,076,000</td>
<td>Urban</td>
<td>133</td>
<td>ICP-OES</td>
<td>46</td>
<td>47</td>
</tr>
<tr>
<td>Mount Isa, Australia</td>
<td>22,000</td>
<td>Mining</td>
<td>60</td>
<td>ICP-AES</td>
<td>346</td>
<td>-</td>
</tr>
<tr>
<td>Miri City, Malaysia</td>
<td>320,000</td>
<td>Urban</td>
<td>37</td>
<td>FAAS</td>
<td>94</td>
<td>12</td>
</tr>
<tr>
<td>Makati, Philippines</td>
<td>510,000</td>
<td>Schools</td>
<td>12</td>
<td>AAS</td>
<td>-</td>
<td>108</td>
</tr>
</tbody>
</table>
2.5. Why soil Pb is still a current issue

Many sources of soil Pb were banned towards the end of the 20th century e.g. leaded-petrol in 1999 and Pb-based paints in 1992 (see Chapter 6). However, legacy Pb contamination in urban soils is still prevalent today due to Pb’s low mobility.

Anthropogenic Pb is predominantly deposited to the topsoil where it migrates slowly to the mineral horizons; therefore, Pb can reside in temperate soils for up to 20,000 years (Kylander et al., 2008; Stille et al., 2011). Initially, Pb exhibits low mobility due to its affinity with OM in the organic (O) horizon, where Pb is strongly bound to ligands or other negative functional groups e.g. carboxylic and humic acids (Gangloff et al., 2014; Chotpantarat et al., 2015). Decomposition of OM causes leaching of Pb to the mineral horizons where it is complexed with Fe and Mn(hydr)oxides e.g. ferrihydrite and birnessite, via inner sphere absorption (Kaste et al., 2006; Schroth et al., 2008). Additionally, clay rich soils have a high cation exchange capacity, so have slower Pb migration rates than sandy soils (Walraven et al., 2013). The above relationships are pH-dependent, as studies on coniferous forest soils have demonstrated that Pb mobility is higher in acidic conditions, due to Pb desorption from complexes and/or the dissolution of Pb minerals e.g. pyromorphite (Stille et al., 2009; Stille et al., 2011). Overall, higher Pb concentrations are observed in the topsoil (<15 cm) compared with deeper horizons (Figure 2.6; Walraven et al., 2014).

Pb mobility in urban soils can be influenced by several factors. Firstly, the addition of de-icing salts to urban roads in the winter causes the dispersion of fine particles and dilution of dissolved organic matter (DOM), which increases Pb mobility in roadside soils (Norrstrom, 2005; Kluge et al., 2014). Lastly, urban soils are often littered with waste materials e.g. construction rubble, and Fe or organic waste,
which can alter the soil properties i.e. increase the pH or provide binding sites, thereby influencing Pb mobility (Norra et al., 2008; Meuser, 2010).

2.6. The compositional nature of geochemical data

Although this research focuses on mapping soil Pb, it is important to acknowledge the compositional nature of geochemical data. Lead is just one element that makes up the soil, with other elements contributing to the soil’s total amount. Subsequently, other geochemical elements present in the soil may influence the total amount of Pb present, and processes which influence Pb may also be found to impact the concentration of other elements (McKinley et al. 2016). Furthermore, maps of a single element may convey patterns or trends that are driven by other elements, and not solely the element being mapped (McKinley et al. 2016). Overall, when a single component in the soil may be influenced by other components this is known as the closure effect or constant sum problem (McKinley et al. 2020). Compositional data analysis (CoDA) techniques can combat this issue, and these are discussed in more detail in Section 10.4.3.

This research focuses on single component analysis for two reasons. Firstly, Pb tends to be anthropogenic in origin (Section 2.4.), and CoDA techniques are more applicable to geogenic elements. Secondly, a map of just soil Pb concentrations is required for determining ‘safe’ soil Pb concentrations and planning future land use(s). For instance, DEFRA’s C4SL are given in mg kg$^{-1}$ for a single element (Section 2.3.). However, there are instances where CoDA analysis could have been undertaken e.g. when examining sources of soil Pb via isotopic analysis (Chapter 7).
2.7. **Conclusions**

This chapter has justified the need to study Pb in urban soils for several reasons. Firstly, as a neurotoxin Pb poses a potential health threat, especially to younger children. Furthermore, even low blood Pb levels can cause neurological damage and a range of other health impacts.

Secondly, soil ingestion and/or inhalation is an important exposure pathway for Pb. Soil Pb can enter the bloodstream from a range of potential routes including intentional ingestion i.e. pica, unintended ingestion of soil adhered to hands or food (i.e. via urban agriculture), and inhalation of resuspended legacy soil Pb. However, not all ingested Pb will be bioaccessible.

The potential health risk posed by soil Pb has caused many governments to quantify the maximum ‘safe’ level of Pb in the soil. In the UK, C4SLs have most recently been adopted, and determine Pb thresholds for different land uses, above which soil Pb may pose a potential threat to human health. These C4SLs, in particular the 200 mg kg\(^{-1}\) threshold for garden soils, will be used in later Chapters (6 and 8) to map soil Pb risk and identify locations at high-risk from high soil Pb concentrations.

Soil Pb concentrations can be elevated in all types of soil around the world. However, urban soils are of a particular concern with regards to Pb, as Pb concentrations in urban soils are more often elevated above background concentrations due to anthropogenic pollution. Additionally, the greater proportion of the population inhabiting urban areas increases the potential risk from soil Pb.

Despite most of the sources of Pb to the soil no longer occurring, soil Pb is still a relevant issue. This is because the legacy of Pb contamination can remain in the topsoil for many years, depending on soil properties.

Lastly, geochemical data is compositional as it generates a constant sum. However, the analysis undertaken in this study focuses on just soil Pb as it generally anthropogenically sourced and single element maps are required for future users.

Overall, the risk from soil Pb needs to be characterised in urban areas, in order to mitigate the potential health impacts posed by high concentrations of soil Pb. To achieve this goal, dataset(s) containing soil Pb concentrations from urban areas will be required. This is discussed in the next chapter.
References - Chapters 1 and 2


residential soil Pb concentrations in Baltimore, Maryland, USA: understanding the variability within cities. *Environmental Geochemistry and Health*, 35 (4), 495–510.


Stille, P., Pierret, M.-C., Steinmann, M., Chabaux, F., Boutin, R., Aubert, D., Pourcelot, L., & Morvan, G., (2009). Impact of atmospheric deposition, biogeochemical cycling and water-mineral interaction on REE fractionation in acidic surface soils and soil water (the Strengbach case). *Chemical Geology*, 264, 173-186


lead-rich particles from a lead recycling plant. Environmental Science and Technology, 45 (18), 7888–7895.

Voutsa, D., & Samara, C. (2002). Labile and bioaccessible fractions of heavy metals in the airborne particulate matter from urban and industrial areas. Atmospheric Environment, 36 (22), 3583–3590.


Chapter 3. Measuring total soil Pb concentrations

Case study datasets with soil Pb concentrations were acquired to explore the spatial variability and controls on soil Pb concentrations (Chapters 5 to 7), and to predict soil Pb values at unsampled locations (Chapters 8 and 9). Greater Glasgow was selected as the study area, due to accessible pre-existing data and the potential to acquire further data. This chapter describes the collection methodology of three datasets measuring total soil Pb concentrations at different scales: the city scale (Greater Glasgow) which was based on an existing dataset, and two neighbourhood scales (Paisley and Bishopbriggs) which were collected for the present study. Organic matter content and pH were also measured in each sample.

3.1. City scale measurements

The BGS has collected soil and other environmental samples throughout the United Kingdom (UK) as part of their G-BASE programme since the 1960s. The G-BASE study initially focused on rural areas, but with growing interest in development on brownfield sites, it was expanded to include urban soil sampling from the 1990s (Flight and Scheib, 2011). To date, the G-BASE study has surveyed 25 urban centres, including Glasgow (Fordyce et al., 2017).

In urban G-BASE surveys samples were collected using a systematic grid. Each British National Grid (BNG) kilometre square was split into four 500 m² squares, in which a sample was collected as close as possible to the square centre (Figure 3.1; Fordyce et al., 2017; Johnston et al., 2005). Each sample was composed of five sub-samples, collected from the corners and centre of a 20 m² sampling area (Figure 3.1; Johnston et al., 2005).

Figure 3.1: Diagram of the G-BASE urban sampling strategy for 1 km BNG square (thick outer box), with black dots representing the location of sub-samples. Red distances show inter-sample spacing (not to scale). Adapted from Johnston et al., 2005.
Topsoil samples were obtained using a handheld Dutch auger to a depth of 20 cm, after removing surface vegetation and leaf litter (Johnston et al., 2005). At each sample location the BNG coordinates of the centre point were recorded using a geographical positioning system (GPS) (Johnston et al., 2005). Additional information was collected, including soil colour, soil texture, clast lithology, site location, geology, land use, and possible contaminants (Fordyce et al., 2017).

After collection, each sample was analysed for pH, OM content, and total concentrations of approximately 50 substances including potentially harmful elements (PHE). The samples were prepared by air-drying then oven-drying (at < 30 °C), and sieving to <2 mm (Fordyce et al., 2017). The sieved soil was then homogenised and ball milled to <53 µm, so that total concentrations of inorganic chemical elements, including Pb, could be measured using X-ray fluorescence spectrometry (XRF) (Fordyce et al., 2017). The pH of the sieved soil was determined by mixing the soil with calcium chloride (CaCl₂), the pH of the resulting slurry was measured using a pH electrode (Fordyce et al., 2017). The OM% was established by loss on ignition (LOI) i.e. heating the sieved soil in a furnace at 450 °C for at least four hours, then calculating the weight difference between the oven- and furnace-dried soil (Fordyce et al., 2017).

The G-BASE survey collected and analysed 2,333 urban and 1,610 rural topsoil samples across the Clyde Basin area (Fordyce et al., 2017). Of these, 1,285 urban topsoil chemical results, covering Greater Glasgow (approximately 320 km²) were selected for inclusion in the present study (Figure 3.2). All urban samples were collected between 2001 and 2002 (Fordyce et al., 2017).
Figure 3.2: Map depicting G-BASE sample locations in Greater Glasgow used for the present study. Insert map shows the position of the study area (in red) in relation to Scotland.
3.2. **Neighbourhood scale measurements**

3.2.1. **Study area selection**

In addition to the city-wide G-BASE dataset, two neighbourhoods in Greater Glasgow were selected to investigate soil Pb concentrations at a finer scale. This helped understand the controls on soil Pb contents. To explore the spatial variability of soil Pb values, ideally the neighbourhood datasets needed to have a large range of soil Pb concentrations. Therefore, high soil Pb outliers were identified in the G-BASE dataset by following the methodology outlined in Rawlins *et al.* (2005). Briefly, the standardised kriging error was calculated from cross-validation of the robust exponential variogram, and a point was classed as an outlier if its standardised kriging error fell above the upper inner fence, according to Tukey's criterion (i.e. $3^{rd}$ quartile+1.5*interquartile range (IQR); Tukey, 1977). This identified 16 high outliers in the Glasgow G-BASE dataset. These were examined in more detail, in particular in relation to the soil Pb concentrations of neighbouring points. The covariate distributions in potential study areas were also considered in order to subsequently explore the controls on soil Pb levels (Chapter 6). Specifically, a range of building ages and historic industries were sought after, as there is limited research investigating their influence on soil Pb concentrations at the fine scale.

Based on these criteria, two neighbourhood study areas were selected as follows.
3.2.1.1. Paisley

Firstly, Ferguslie or Linwood in Paisley was selected due to its large range in soil Pb concentrations (55 to 1,458 mg kg\(^{-1}\)) and variety of historic industry (Figure 3.3). Paisley is the largest town in Renfrewshire, located west of Glasgow on the banks of the White Cart Water which feeds into the River Clyde (Paisley Scotland, 2022). The study area is situated in west Paisley and follows Barskiven Road to encompass part of Linwood retail park in the west and Ferguslie Park to the east (Figure 3.3). The southwest of the study area was originally the Linwood Steel Factory where railway rolling stock was manufactured (Figure 3.4; Canmore, n.d.). However, in 1963 the factory was bought by the Rootes Group and expanded to manufacturing cars, most notably Hillman Imp cars (Kennouche, 2015). In 1976 the plant was sold to Peugeot-Citroen but was soon demolished and developed into a retail park, as it can be found today (Kennouche, 2015). There are also two dismantled railways at the site, the 1858 Linwood Branch Line and the later 1912 Paisley-Barrhead Line.

According to historical maps (Crown Copyright, 2021), building ages in Paisley are rather homogeneous. Only one building in Ferguslie Park was constructed before 1920, with the majority of buildings erected post-war but pre-1950s, as this was when Ferguslie Park council housing estate was developed (FP HA, 2019). There are also some newer (1992) buildings to the west of the study area, built on old industrial land. The study area also includes a mix of land uses, with 46% of the area classified as private gardens and 33% as semi-natural.

Figure 3.4: Historic photograph of Linwood Steel works taken on 09/02/1950. The photograph faces north, so the approximate location of the west part of the study area has been outlined in red. © HES (Aerofilms Collection) from Canmore (n.d.).
Figure 3.3: Map of the study area in Paisley showing G-BASE samples with their measured soil Pb concentrations, and digitised historic industry. The insert depicts the area covered by the G-BASE samples selected for this study (dark grey) in relation to the study area in Paisley (red box).
3.2.1.2. Bishopbriggs

Secondly, Bishopbriggs in northeast Glasgow was selected as G-BASE samples from this area show a large range of soil Pb levels (175 to 1,019 mg kg⁻¹, Figure 3.5). Originally called Bishop’s Riggs, this area ~6 km north of Glasgow city centre developed from a small village to a residential suburb and commuter town of Glasgow (GB Historical, 2022). Subsequently, this area contains a range of building ages (Figures 3.5 and 3.6), which were determined using historic maps (Crown Copyright, 2021). For instance prosperous Victorian (pre-1860s) villas can be seen along Springfield Road (Figure 3.6) and slightly later tenements (built around 1900) are evident around Ruskin Square (Figure 3.5). Inter-war council houses (built 1910s - 1930s) are also present along Springfield Square and Emerson Road (Figure 3.5). These are municipal two-story cottages with spacious front and rear gardens, which were popular in Glasgow in the inter-war years (Maver, 2000). Most recently, the area around Woodfield Avenue was developed with modern detached and semi-detached houses (post-1970s; Figure 3.5).

Historical industry in northeast Glasgow was limited to early (19th century) sandstone quarrying and slightly later (20th century) coal mining. This contrasts with the wide range and high density of historic industry in Paisley. However, the study area has varied land uses, e.g. domestic gardens, two parks or public open space, commercial property, and some woodland.

![Springfield Road, Bishopbriggs](image)

**Figure 3.6:** Historic photograph of Springfield Road taken in the early 1900s, looking west. Photograph location is shown in Figure 3.5. From Bishopbriggs History Group (2021).
Figure 3.5: Map of the study area in Bishopbriggs showing G-BASE samples with their measured soil Pb concentrations and digitised building age. The insert depicts the area covered by the G-BASE samples selected for this study (dark grey) and the study area in Bishopbriggs (red box). The star shows the location of the photograph in Figure 3.6.
3.2.2. Sampling methodology

In order to understand the fine scale controls and variability of soil Pb within each study area samples were collected using a grid pattern. This also corresponds to the G-BASE sampling methodology (Johnson et al., 2005). Each study area was split into a grid of 100 cells each of 42 m$^2$, with the centre of each grid as the sample location (Figure 3.7). In Paisley the study area was 700 m by 250 m (0.175 km$^2$). The study area in Bishopbriggs was extended to ensure that the G-BASE high outlier was completely encompassed (study area = 742 x 292 m = 0.22 km$^2$). The sample locations were adjusted manually, using aerial photography, so that they fell on greenspace, and the address and coordinates of each sample were determined. During sampling, access issues sometimes required slight adjustments of sample location, but the average distance between the predetermined grid location and actual sample location was small at 6 m.

To assess the even finer scale range in soil Pb content, ten grid samples were selected at random from each study area, and an additional sample was collected 10 m from the original sample. This follows Cattle et al.’s (2002) recommendation to include extra pairs of samples closer together, which will help any future variograms accurately identify variance at shorter lags. The exact location of the additional sample depended on access and available greenspace. Furthermore, five samples were selected at random from each study area, and duplicate samples were collected to check sampling precision. Additionally, the recorded location of any G-BASE sample points in the study area were sampled. Overall 260 samples were collected (Paisley = 100, Bishopbriggs = 126, G-BASE repeat = 4, duplicates = 10, 10 m extra samples = 20; Figure 3.7).

In Bishopbriggs, samples were collected over five days from November 4th to December 2nd, 2017, and in Paisley samples were gathered over four days from April 4th to May 2nd, 2018. Prior to sampling, permission was granted from the relevant local authority for samples collected from public land. Individual owner(s) gave verbal consent before any samples were collected from private land.
Figure 3.7: Maps showing sample locations in both study areas, Paisley and Bishopbriggs, coloured by type of sample. The insert depicts the area covered by the G-BASE samples selected for this study (dark grey) in relation to the study areas (red boxes).
The sampling methodology aimed to accurately represent soil Pb contents in each study area. Firstly, three composite samples from the corners of a triangle (edges of 1 m; Figure 3.8) were aggregated to ensure representative samples were collected. This is a smaller aggregation area than the G-BASE survey, where a 20 m$^2$ area was used for composite sampling (Johnson et al., 2005). However, analysis of the G-BASE dataset indicates the larger composite area may have masked the fine scale influence of roads and paint (Chapter 6). The coordinates of the centre of the triangle (BNG) were recorded with a handheld GPS (Garmin eTrex 20x, accuracy ±4 m). Photographs and field notes were also recorded. Samples were collected in a random order but analysed numerically to assess for analytical drift (Plant et al., 1975).

A handheld Dutch auger was used to sample topsoil to depth of 20 cm, after surface vegetation and leaf litter were removed. The auger head was cleaned with deionised water between sample locations to prevent cross contamination. This is the same methodology employed during the G-BASE project (Johnson et al., 2005). Collected samples were placed in a new, labelled plastic bag, which was double bagged and labelled to prevent cross-contamination and sample mix-ups.

![Figure 3.8: Photographs of sampling methodology; with 1 m triangle sampling area shown, with labelled sample bag, and GPS recording central coordinates.](image)

The samples were returned to the laboratory for chemical analysis. This was carried out in two phases. The aim of the initial analytical round was to determine the total Pb concentrations, OM percentage and pH of the 260 samples collected from Paisley and Bishopbriggs. The methodologies for these analyses are described here. Subsequently, Pb-isotope analysis was conducted on a sub-set of soil samples to examine Pb sources, as explained in Chapter 7.
3.2.3. Initial laboratory analysis

Samples were prepared for analysis firstly by air drying until a constant weight was reached. The samples were then sieved to <2 mm to homogenise the three cores and remove any material that was not soil. The sieved samples were stored in labelled, double plastic bags, at room temperature, in a dark location. A series of quality control checks were performed to confirm the accuracy and precision of the results.

3.2.3.1. Total Pb concentration

To determine total Pb concentrations the USEPA method 3052 (i.e. microwave assisted acid digest; USEPA 1996) was used to dissolve the soil into solution, which was then analysed using inductively coupled plasma optical emission spectroscopy (ICP-OES). This method was selected as it is proven to produce accurate results when measuring soil Pb concentrations (e.g. Yafa and Farmer, 2005; Alagic et al., 2018).

To ensure homogeneity and complete digestion, the samples were coned and quartered to obtain a ~5 g subsample. This subsample was ball-milled (using a RETSCH Mixer Mill 200) to reduce the particle size to <53 μm. From this material a further 0.25 g (±0.00047 g) subsample was taken for analysis. This subsample was oven dried at 105 °C until a constant weight was achieved (approximately 24 hours), which removed any residual moisture. Next, the sample was ashed in a muffle furnace at 450 °C for four hours to remove any organics. The remaining material was mixed with 9 ml of concentrated (69%) Aristar™ nitric acid (HNO₃) and 1 ml of 40% Aristar™ hydrofluoric acid (HF) in a CEM MARS5 microwave for 10 minutes at 180 °C, with >20 minutes for cooling (USEPA method 3052, USEPA 1996). The digested material was evaporated on a hotplate to 1-2 ml, and then diluted to 25 ml using 2% Aristar™ HNO₃. Finally, the samples were filtered into Sterilin™ tubes through Whatman™ grade 540, hardened, ashless filter papers and stored in a cool (4 °C), dark environment until analysis.

To reduce cross-contamination all glassware was cleaned in a hot acid bath (Analar™ 5 M HNO₃) for four hours, rinsed with deionised water, then placed in a hot deionised water bath for a further four hours. Subsequently, glassware was rinsed with deionised water and left to dry in a clean drying cupboard (60 °C). Plasticware was cleaned in a cold acid bath (Analar™ 5 M HNO₃) for 24 hours, then rinsed with deionised water, and left to dry in a clean drying cabinet. Microwave
vessels were cleaned with 10 ml of concentrated (40%) Analar™ HNO₃, using the microwave cleaning programme. Any other equipment was cleaned with methanol between samples. Two reagent blanks were also included per microwave run to determine cross-contamination. All samples were analysed in duplicate to assess precision, and two certified reference materials were included per microwave run to measure accuracy.

Total Pb concentrations in the digested solution were determined by ICP-OES. During ICP-OES analysis the liquid sample is pumped (via a peristaltic pump) into a nebulizer, where argon gas converts the sample to a mist (Wang, 2004). The spray chamber is then used to separate the fine sample droplets (i.e. <10 µm) from the coarser aerosol droplets, which are emitted as waste (Wang, 2004; Thomas, 2004). The fine droplets enter the plasma torch which excites electrons from their ground state to a higher energy level (Thomas, 2004). When these electrons fall back to their ground state, they emit photons or electromagnetic radiation, which has specific wavelengths depending on the element (Thomas, 2004). Concentrations of the studied element are obtained by comparing the observed intensity with a calibration curve, which is generated from a blank and a range of standard solutions (Thomas, 2004).

The G-BASE study used XRF to measure total Pb concentrations (Fordyce et al., 2017); however ICP-OES was selected in this study for several reasons. Firstly, although soil digestions can be costly and time consuming, they move the Pb into solution, meaning Pb isotope ratios can also be measured. Subsequently, ICP-OES is more commonly used than XRF in studies measuring Pb isotopes along with soil Pb (Broadway et al., 2010; Rodríguez-Seijo et al., 2015). ICP-OES also has a wide analysis range (4-9 orders of magnitude, depending on calibration standards), and it examines multiple spectral lines so one wavelength should have little interferences (Wang, 2004). Both XRF and ICP-OES can analyse multiple elements if the correct standards are used.

To confirm that results generated by ICP-OES and XRF would be comparable, a small exploratory study was undertaken. Five G-BASE samples were reanalysed in triplicate using the acid-assisted microwave digestion and ICP-OES methodology, detailed above. The measured Pb concentrations for the same sample, but using a different analysis method, were compared. The overall Pearson’s R² correlation coefficient was 0.97 (p<0.01, for log transformed data), indicating a strong
agreement between the methods. Other studies have also observed a close agreement in Pb concentrations analysed by XRF and ICP-OES (Margui et al., 2005; Barsby et al., 2012).

This study used a Perkin Elmer Optima 5800 DV ICP-OES, measuring the wavelength 220.35 nm for Pb. This wavelength was selected as it has limited Fe interference, compared with other Pb wavelengths e.g. 217 nm. The instrument parameters are outlined in Table 3.1. Before analysis the instrument was calibrated using at least six ICP multi-element standard solutions VI (M4; Merck), ranging from 0-100 parts per million (ppm) Pb.

<table>
<thead>
<tr>
<th>Table 3.1: Operating parameters used by the ICP-OES during analysis.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Parameter</strong></td>
</tr>
<tr>
<td>Radio frequency power</td>
</tr>
<tr>
<td>Nebuliser</td>
</tr>
<tr>
<td>Spray chamber</td>
</tr>
<tr>
<td>Argon flow rates (L min⁻¹)</td>
</tr>
<tr>
<td>Peristaltic pump flow</td>
</tr>
<tr>
<td>Viewing mode</td>
</tr>
</tbody>
</table>

During analysis external standard solutions (1 ppm and 10 ppm Pb) were analysed to verify the calibration. Additionally, a drift check (i.e. 1 ppm M4 standard reanalysed) was run every ten samples to ensure the instrument was accurately measuring Pb throughout the analysis. Soil Pb concentrations were determined by subtracting the average blank concentration ($\mu_{blank}$) from the measured concentration ($C$), which was multiplied by the dilution volume ($V$ in litres) divided by the ashed sample weight ($W_{tA}$, kilograms):

$$Pb \text{ (mg kg}^{-1}) = (C - \mu_{blank}) \ast \frac{V \text{ (l)}}{W_{tA} \text{ (kg)}}.$$  \hspace{1cm} (3.1)

These results were post-processed to find the most appropriate calibration range, and to adjust for machine drift (see Appendix 3.1).

3.2.3.2. pH

To measure the pH of the samples 5 g of <2 mm air-dried soil was combined with 25 ml of 0.01 molar CaCl$_2$ (Baxter, 2018). This was mixed in an end-over-end shaker for 20 minutes at 30 rpm. The pH of the resulting slurry was measured using a pH probe (Jenway 3505 bench pH meter). Each run included approximately seven samples and one duplicate sample. Before each run the pH probe was calibrated.
using pH 4 and 7 buffer solutions (Fischer Scientific™). Between samples the probe was cleaned with deionized water.

3.2.3.3. Organic matter
The OM content of the samples were calculated in duplicate using LOI. This determines the OM% using the percentage difference between the weight of the oven dried soil ($W_{t_{OD}}$) and the weight of the ashed soil ($W_{tA}$; Kenkel, 2003):

\[
    \text{OM\%} = \frac{W_{t_{OD}} - W_{tA}}{W_{t_{OD}}} \times 100.
\]

3.2.3.4. Quality control checks
Contamination from laboratory processing and analysis can cause serious inaccuracies in geochemical measurements (Plant et al., 1975). Cross-contamination was assessed using reagent blanks i.e. no soil material only chemicals, which were included as two samples per microwave-assisted acid digestion run, with soil Pb concentrations (ppm) determined during ICP-OES analysis. The resulting Pb values for the reagent blanks are low, with a maximum overall mean concentration of 0.012 ppm Pb (Table 3.2). This indicates no cross-contamination occurred during analysis.

| Table 3.2: Concentrations of reagent blanks (ppm) for Bishopbriggs, Paisley, and all samples. 1standard deviation, 2limit of detection. *0 substituted for negative results. |
|---|---|---|---|---|---|---|
| Count | S.D. | Mean | Min. | Max. | LOD |
| Bishopbriggs | 56 | 0.028 | 0.017 | 0* | 0.12 | 0.11 |
| Paisley | 48 | 0.012 | 0.0065 | 0* | 0.06 | 0.05 |
| All | 104 | 0.022 | 0.012 | 0* | 0.12 | 0.08 |

During ICP-OES analysis, the smallest concentration of Pb that can be accurately measured is defined using the limit of detection (LOD). The LOD was calculated using the blank method i.e. mean of all blanks plus 3.3 standard deviations of the blanks (Armbruster and Pry, 2008):

\[
    LOD = \mu_{\text{blank}} + 3.3 \sigma_{\text{blank}}.
\]

The overall LOD is 0.08 ppm (also 0.08 mg kg⁻¹; Table 3.2), which is acceptable as the sample with lowest Pb concentration contained more Pb (0.11 ppm). It is also comparable with other studies using ICP-OES to measure soil Pb, where the LOD was 0.02 ppm to 0.05 ppm (Witsche and Wolfgang, 2020).

To assess the digestion and overall analysis accuracy, two reference samples were included per digestion run. For Bishopbriggs certified reference material (CRM)
light sandy soil 7002 containing 58.9±4.9 mg kg⁻¹ of Pb was used, and CRM silty clay loam 7003 was used for Paisley, which contains 33.5±2.4 mg kg⁻¹ of Pb (Analytika, Prague, Czech Republic). As two different CRMs were used, the recovery rate was calculated for each area. This was determined using the difference between the CRM mean Pb concentration ($P_{\text{Cert}}$) and the mean measured Pb levels ($P_{\text{Meas}}$):

$$\text{Recovery} = 100 - \left( \frac{P_{\text{Cert}} - P_{\text{Meas}}}{P_{\text{Cert}}} \times 100 \right) . \quad (3.4)$$

The overall recovery rate is very good at 98% (Table 3.3). However, the average recovery rate in Bishopbriggs is excellent at 100%, compared with Paisley where it is slightly lower (mean = 95%; Table 3.3). The lower recovery rate in Paisley may be due to the Paisley CRM containing lower Pb levels (~50% less than Bishopbriggs CRM); therefore, any Pb lost or not completely digested will have a bigger impact on the recovery rate.

| **Table 3.3:** Certified reference material values for Bishopbriggs, Paisley, and all samples. Difference from certified value (mg kg⁻¹) is given, with the recovery rate (Equation 3.4). |
|-----------------|-----|-----|-----|-----|-----|
| **Bishopbriggs** | **Mean** | **S.D.** | **Count** | **Min.** | **Max.** |
| Difference      | 6.83 | 5.24 | 53  | 0.08 | 22.34 |
| Recovery %      | 100.32 | 14.7 |     | 70.06 | 137.94 |
| **Paisley**     | **Mean** | **S.D.** | **Count** | **Min.** | **Max.** |
| Difference      | 3.93 | 3.29 | 46  | 0.02 | 13.26 |
| Recovery %      | 95.45 | 14.08 |   | 60.43 | 136.43 |

| **Table 3.4:** Comparison between measured and certified values for external check (M4) during Bishopbriggs, Paisley, and all ICP-OES runs. Difference from certified value (ppm) is given, along with the recovery rate (Equation 3.4). |
|-----------------|-----|-----|-----|-----|-----|
| **Bishopbriggs** | **Mean** | **S.D.** | **Count** | **Min.** | **Max.** |
| Difference      | 0.60 | 0.32 | 9   | 0.079 | 1.008 |
| Recovery %      | 98.96 | 7.52 |   | 91.02 | 109.97 |
| **Paisley**     | **Mean** | **S.D.** | **Count** | **Min.** | **Max.** |
| Difference      | 0.33 | 0.42 | 8   | 0.024 | 1.026 |
| Recovery %      | 100.99 | 6.92 |   | 90.62 | 110.26 |
| **All**         | **Mean** | **S.D.** | **Count** | **Min.** | **Max.** |
| Difference      | 0.47 | 0.37 | 17  | 0.024 | 1.026 |
| Recovery %      | **99.98** | **7.22** |   | 90.62 | 110.26 |

To assess the accuracy of the ICP-OES instrument an external standard Pb solution (ICP multi-element standard solution VI (Merck; M4)) was used. The M4 solution was diluted to 1 ppm and 10 ppm Pb, to represent the range in Pb values present in the samples. These standards were analysed in duplicate after calibration but before analysis. As with the CRM, recovery rates give a good indication of instrument accuracy. The resulting recovery rates are very good as they are close to 100% with
a <10% range (Table 3.4). This shows the ICP-OES was accurately calibrated for all runs. Furthermore, the calibration curves display an average correlation coefficient of 0.9998 (Appendix 3.1; higher correlation coefficient for lower ranges of Pb standards). This signifies the instrument was correctly calibrated before analysis.

To ensure the ICP-OES remained accurately calibrated throughout the run the 1ppm M4 standard was reanalysed every ten samples. In several of the runs the measured value tended to decrease continuously throughout the run, indicating lower accuracy for samples analysed later in the run (Figure 3.9). This issue could have been solved by running internal yttrium standards (e.g. Zachariadis and Vogiatzis, 2010). This was tested in one run (BB2), but the results were not significantly improved and the process was more time consuming and subject to human error. Instead the decrease in the 1 ppm standard with increasing sample number was modelled, and a best fit equation was applied to all blank corrected Pb concentrations in solution (Marcos and Hill, 2000; Table 3.9 Appendix 3.1).

Figure 3.9: Lead concentration (ppm) of 1ppm M4 standard solution over time, with sample analysis number, for the 1st Bishopbriggs ICP-
Laboratory precision was determined by comparing the 250 laboratory duplicate results. Specifically the standard deviation and relative standard deviation (RSD = SD/Mean*100) were calculated for every sample. If the RSD was >10% (n = 20 in Bishopbriggs, = 19 in Paisley), then a triplicate sample was analysed and the mean concentration of soil Pb was determined from the three samples. Running triplicate analysis of select samples substantially reduced the average RSD in both Paisley and Bishopbriggs to 5% (overall mean SD = 21 mg kg⁻¹; Table 3.5). Although there are still some outliers (maximum RSD = 43%). In Paisley the decrease in mean RSD was greater (28% decrease; Table 3.5) than the Bishopbriggs decrease (1%) due to the reduction of outliers. The average RSD of laboratory duplicates in this study (5%) is similar to the RSD% determined by Minkkinen (1986) of 8%.

Field precision was assessed via differences in soil Pb levels between the ten field duplicates collected during sampling (Table 3.6). The mean absolute difference (Mean.AD) was calculated to quantify precision as it is more robust to outliers that other precision metrics because it does not square differences (Hyslop and White, 2009). The Mean.AD uses the difference between duplicates ($D_i$) to determine precision, where the individual sample’s mean difference ($S_1$ and $S_2$) is divided by the overall mean ($\bar{S}_i$) (Hyslop and White, 2009):

$$D_i = \frac{(S_1 - S_2)}{\sqrt{2}/\bar{S}_i}.$$  \hspace{1cm} (3.5)

The Mean.AD precision is then determined by:

$$\text{Mean.AD} = \sqrt{\frac{\pi}{2}} \frac{1}{n} \sum_{i=1}^{n} |D_i| \times 100\%.$$ \hspace{1cm} (3.6)
Table 3.6: Field duplicate results, with sample number and area (BB=Bishopbriggs, P=Paisley) and mean soil Pb concentrations (mg kg⁻¹) for each duplicate sample. The relative difference (mg kg⁻¹) and Mean.AD are shown for each sample. Red cells indicate a large (>10%) Mean.AD, and green a small Mean.AD (<10%).

<table>
<thead>
<tr>
<th>Sample</th>
<th>Sample 1 mean Pb (kg mg⁻¹)</th>
<th>Sample 2 mean Pb (kg mg⁻¹)</th>
<th>Difference (mg kg⁻¹)</th>
<th>Mean.AD (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB39</td>
<td>117.33</td>
<td>131.55</td>
<td>14.22</td>
<td>14.70</td>
</tr>
<tr>
<td>BB55</td>
<td>340.49</td>
<td>156.39</td>
<td>184.09</td>
<td>32.99</td>
</tr>
<tr>
<td>BB89</td>
<td>158.20</td>
<td>157.85</td>
<td>0.36</td>
<td>0.06</td>
</tr>
<tr>
<td>BB114</td>
<td>1122.27</td>
<td>955.52</td>
<td>166.75</td>
<td>7.10</td>
</tr>
<tr>
<td>BB118</td>
<td>278.49</td>
<td>184.65</td>
<td>93.84</td>
<td>18.07</td>
</tr>
<tr>
<td>P41</td>
<td>94.63</td>
<td>95.90</td>
<td>1.27</td>
<td>0.39</td>
</tr>
<tr>
<td>P57</td>
<td>130.59</td>
<td>127.74</td>
<td>2.85</td>
<td>2.95</td>
</tr>
<tr>
<td>P59</td>
<td>889.11</td>
<td>846.62</td>
<td>42.49</td>
<td>2.38</td>
</tr>
<tr>
<td>P68</td>
<td>127.52</td>
<td>129.72</td>
<td>2.20</td>
<td>0.52</td>
</tr>
<tr>
<td>P84</td>
<td>523.40</td>
<td>741.46</td>
<td>218.06</td>
<td>15.07</td>
</tr>
</tbody>
</table>

Overall Mean.AD 12.45

The difference in soil Pb levels between the field duplicates is inconsistent, as six samples are in very close agreement (<10% Mean.AD), whereas four samples show a large variation in soil Pb levels (>10% Mean.AD; Table 3.6). This could be due to some samples containing small fragments of Pb that were not present in the other duplicate, which could also cause the high RSD between a few laboratory duplicates (Table 3.5). However, Hyslop and White’s (2009) analysis of the IMPROVE dataset, which measured Pb concentrations in particulate matter in duplicate at 715 sites, calculated the Mean.AD precision of 10%. This is comparable to the overall Mean.AD in this study of 12.5% (Table 3.6), although the present study is based on fewer duplicate pairs and studies soil not air quality.

Samples with a large difference between duplicates (i.e. >10% Mean.AD for field duplicates or RSD>10% for laboratory duplicates, n = 24), were extracted, and their covariate values compared with the remaining dataset. The samples with a large difference in duplicate soil Pb values tended to be closer to buildings (mean distance = 9.7 m) compared with the whole dataset (mean distance = 18.2 m). Furthermore, 25% of dissimilar samples were close to newer (post-1970) buildings, which is higher than the proportion of the whole dataset close to new buildings (17%). This could indicate locations where the soil has been reworked to accommodate new buildings, which may alter the distribution of Pb in the soil.
Lastly, the difference between soil Pb measurements from the same location in the G-BASE survey and Paisley or Bishopbriggs in this study were compared (Table 3.7). The differences were inconsistent, with some nearby samples showing a 94% difference in soil Pb concentrations, and other samples with no (0%) differences (Table 3.7). These discrepancies may be due to differences in sampling methodologies i.e. the G-BASE study aggregated soil from a larger area compared with this study. Alternatively, land use may have changed in the 15 years between sample collections. Lastly, microscale variations are known to occur in soil Pb, and could influence soil Pb concentrations, this is discussed in more detail in Chapter 5.

Overall, the quality control checks indicate that the soil Pb concentrations are accurate and precise enough to explore soil Pb variability, controls, and predictive modelling. For example, a difference of 21 mg kg\(^{-1}\) (average SD for laboratory duplicates), will be unlikely to change if a sample has high (≥ 200 mg kg\(^{-1}\)) Pb or low Pb (< 200 mg kg\(^{-1}\)). Furthermore, the heterogenous nature of Pb in urban soils has been noted in other studies and these results are within the bounds of similar research.
<table>
<thead>
<tr>
<th>Location</th>
<th>G-BASE ID</th>
<th>Bishopbriggs NW 601345</th>
<th>Bishopbriggs NE 610370</th>
<th>Paisley NW 610452</th>
<th>Paisley NE 610405</th>
</tr>
</thead>
<tbody>
<tr>
<td>Map</td>
<td></td>
<td><img src="image1" alt="Map Image" /></td>
<td><img src="image2" alt="Map Image" /></td>
<td><img src="image3" alt="Map Image" /></td>
<td><img src="image4" alt="Map Image" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>G-BASE soil Pb (mg kg⁻¹)</th>
<th>175 mg kg⁻¹</th>
<th>1019 mg kg⁻¹</th>
<th>1458 mg kg⁻¹</th>
<th>55 mg kg⁻¹</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>This study</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Nearest sample ID</td>
<td>48</td>
<td>37</td>
<td>35</td>
<td>4</td>
</tr>
<tr>
<td>Pb conc. (mg kg⁻¹)</td>
<td>100 mg kg⁻¹</td>
<td>66 mg kg⁻¹</td>
<td>1156 mg kg⁻¹</td>
<td>119.8 mg kg⁻¹</td>
</tr>
<tr>
<td>Sample distance (m)</td>
<td>8 m</td>
<td>6 m</td>
<td>7 m</td>
<td>5 m</td>
</tr>
<tr>
<td>Difference (abs, mg kg⁻¹)</td>
<td>75 mg kg⁻¹</td>
<td>953 mg kg⁻¹</td>
<td>303 mg kg⁻¹</td>
<td>64 mg kg⁻¹</td>
</tr>
<tr>
<td>Difference (%)</td>
<td>43%</td>
<td>94%</td>
<td>21%</td>
<td>116%</td>
</tr>
<tr>
<td>2nd nearest sample ID</td>
<td>101</td>
<td>76</td>
<td>38</td>
<td>49</td>
</tr>
<tr>
<td>Pb conc. (mg kg⁻¹)</td>
<td>86 mg kg⁻¹</td>
<td>242 mg kg⁻¹</td>
<td>928 mg kg⁻¹</td>
<td>55 mg kg⁻¹</td>
</tr>
<tr>
<td>Sample distance (m)</td>
<td>32 m</td>
<td>36 m</td>
<td>11 m</td>
<td>32 m</td>
</tr>
<tr>
<td>Difference (abs, mg kg⁻¹)</td>
<td>88 mg kg⁻¹</td>
<td>777 mg kg⁻¹</td>
<td>530 mg kg⁻¹</td>
<td>0 mg kg⁻¹</td>
</tr>
<tr>
<td>Difference (%)</td>
<td>51%</td>
<td>76%</td>
<td>36%</td>
<td>0%</td>
</tr>
</tbody>
</table>
References - Chapter 3


Analytika, Prague, Czech Republic. CRM soil. Accessed online on 24/01/2022 from: www.onlinecas.com/Etalons-Standards/standardAnalytika/CRMSOILdonneuk.html


Digimap/Crown Copyright and Landmark Information Group Limited (2021). All rights reserved. *Historic maps from 1860-2000* Downloaded from: https://digimap.edina.ac.uk/roam/download/historic


### Appendix 3.1.

**Table 3.9: Correlation coefficient of best-fit line from standard solutions for low (0-5ppm), medium (0-10ppm) and high (0-20ppm) M4 standards. Overall mean = 0.999625. Standard range selected depended on range of Pb in samples.**

<table>
<thead>
<tr>
<th>Run</th>
<th>0-5ppm correlation coefficient</th>
<th>0-10ppm correlation coefficient</th>
<th>0-20ppm correlation coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB1</td>
<td>0.99995</td>
<td>0.999987</td>
<td>0.99901</td>
</tr>
<tr>
<td>BB2</td>
<td>n/a</td>
<td>0.999980</td>
<td>0.999882</td>
</tr>
<tr>
<td>BB3</td>
<td>0.999842</td>
<td>0.999889</td>
<td>0.999541</td>
</tr>
<tr>
<td>BB4</td>
<td>0.999994</td>
<td>0.999815</td>
<td>0.999303</td>
</tr>
<tr>
<td>BB4A</td>
<td>n/a</td>
<td>0.999854</td>
<td>n/a</td>
</tr>
<tr>
<td>BB5</td>
<td>0.999941</td>
<td>0.999949</td>
<td>0.999825</td>
</tr>
<tr>
<td>BB6</td>
<td>0.999942</td>
<td>0.999928</td>
<td>0.999881</td>
</tr>
<tr>
<td>P1</td>
<td>0.999362</td>
<td>0.999853</td>
<td>n/a</td>
</tr>
<tr>
<td>P2</td>
<td>0.99994</td>
<td>n/a</td>
<td>0.99994</td>
</tr>
<tr>
<td>P3</td>
<td>0.999999</td>
<td>n/a</td>
<td>0.99969</td>
</tr>
<tr>
<td>P4</td>
<td>0.99988</td>
<td>n/a</td>
<td>0.99930</td>
</tr>
<tr>
<td>Extra</td>
<td>0.999945</td>
<td>n/a</td>
<td>0.99953</td>
</tr>
<tr>
<td>Mean</td>
<td>0.99896</td>
<td>0.99907</td>
<td>0.99673</td>
</tr>
</tbody>
</table>

**Table 3.10: Drift correction applied to samples from each run. Person’s correlation coefficients and formula are given for the best fit linear and quadratic models from 1ppm drift checks. S = sample number. Equation with the highest $R^2$ values were implemented.**

<table>
<thead>
<tr>
<th>Run</th>
<th>Linear</th>
<th>Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>Equation</td>
</tr>
<tr>
<td>BB1</td>
<td>0.91</td>
<td>$1.037 \times (0.0031\times S)$</td>
</tr>
<tr>
<td>BB2</td>
<td>n/a internal standards used</td>
<td>n/a</td>
</tr>
<tr>
<td>BB3</td>
<td>0.69</td>
<td>$1.055\times (0.0017\times S)$</td>
</tr>
<tr>
<td>BB4</td>
<td>0.81</td>
<td>$0.98\times (0.0085\times S)$</td>
</tr>
<tr>
<td>BB4B</td>
<td>0.99</td>
<td>$13.66\times (0.037\times S)$</td>
</tr>
<tr>
<td>BB5</td>
<td>0.86</td>
<td>$1.0099\times (0.00061\times S)$</td>
</tr>
<tr>
<td>BB6</td>
<td>0.97</td>
<td>$1.015\times (0.0012\times S)$</td>
</tr>
<tr>
<td>P1</td>
<td>0.92</td>
<td>$1.15\times (0.0028\times S)$</td>
</tr>
<tr>
<td>P2</td>
<td>0.99</td>
<td>$1.085\times (0.0021\times S)$</td>
</tr>
<tr>
<td>P3</td>
<td>0.98</td>
<td>$1.016\times (0.0022\times S)$</td>
</tr>
<tr>
<td>P4</td>
<td>0.99</td>
<td>$1.021\times (0.00194\times S)$</td>
</tr>
<tr>
<td>Extra</td>
<td>0.99</td>
<td>$1.064\times (0.0024\times S)$</td>
</tr>
</tbody>
</table>
Chapter 4. Exploratory data analysis

Prior to commencing any statistical analyses the soil Pb concentration from Greater Glasgow and the two neighbourhood datasets were explored. This aimed to determine if the data followed a Normal Gaussian distribution and were isotropic.

In soil geochemistry the data are often not normally distributed but are positively skewed. Therefore, the median is typically preferred over the mean as an average measure as it is less influenced by skew in the datasets (Lark and Lapworth, 2012). Standard geostatistical models require the assumption that the data are normally distributed. This assumption is not appropriate if the distribution of the data is highly skewed (i.e. the distribution is asymmetric with a small number of large values). Along with a Normal distribution, standard spatial analysis and prediction procedures also assumes that the data is isotropic i.e. spatial variation is not influenced by direction (Oliver and Webster, 2015). Therefore, any anisotropy or spatial trends should be identified in the data. To correct for a skewed distribution or anisotropic data a transformed is often applied before analysis. Oliver and Webster (2015) recommend only transforming the data if the skew or anisotropy will influence further analysis, otherwise the original data should be preserved.

4.1. Methodology

All exploratory data and spatial analyses in this chapter were performed in R (Version 4.1.2), with example R script provided in Appendix 4.1. Summary statistics, including the mean, median, standard deviation, median absolute deviation (Med.AD i.e. the median of the absolute difference between all soil Pb values and overall median), minimum, and maximum were calculated for soil Pb concentrations in each dataset.

The skew of each soil Pb dataset was determined using the octile skew. The octile skew (Brys et al., 2003) compares the symmetry in the 1st octile ($O_1$) and 7th octile ($O_7$) about the median ($O_4$):

$$\text{Octile Skew} = \frac{(O_7 - O_4) - (O_4 - O_1)}{O_7 - O_1}. \tag{4.1}$$

The octile skew is a more robust measure of skew than traditional measurements because it is less sensitive to extreme outliers (Lark et al., 2006). If the octile skew is $>0.2$ or $<-0.2$ it indicates a transform is required (Lark and Bishop, 2007).
If required, the soil Pb data was transformed using the Box-Cox family of transformations. A Box-Cox transformation converts a continuous positive variable \( z \) to a new variable \( bc.Z \), using lambda \( \lambda \) (Diggle and Riberio, 2007):

\[
bc.Z = \begin{cases} 
  (z^\lambda - 1)/\lambda : & \lambda \neq 0 \\
  \log_z : & \lambda = 0
\end{cases}.
\]  

An appropriate value of \( \lambda \) will lead to an approximately normally distributed transformed variable. As Equation 4.2 demonstrates, when lambda is zero a standard log transform is computed on the variable \( z \). However, the optimal lambda value is often not 0, so lambda should be estimated as part of the model fitting procedure (Lark et al., 2014). The optimal lambda value is then used to transform the data following the top line in Equation 4.2.

Isotropy in the soil Pb datasets were assessed using linear regression models, which were fit using ordinary least squares (OLS; Lark et al., 2006). The regression models compared soil Pb (Box-Cox transformed) concentrations with the sample Easting or Northing (BNG) and determined the strength and significance of the relationship. Additionally, directional variograms were generated which plot the spatial dependence or semivariance between pairs of points at varying directions from each other e.g. \( 0^\circ, 45^\circ, 90^\circ \) and \( 135^\circ \) (Gringarten and Deutsch, 2001). Isotropic datasets will have similar variograms regardless of direction.

**4.2. Chapter aims**

This chapter aims to determine:

1. The soil Pb concentrations for each datasets.

2. If any of the datasets have a skewed distribution? If so, can this be corrected for?

3. If any of the datasets have an isotopic pattern? If so, can this be corrected for?
4.3. **Results and discussion**

4.3.1. **Soil Pb concentrations**

Summary statistics were calculated for soil Pb levels in each study area (Table 4.1), along with frequency histograms, to assess the data distribution (Figure 4.1).

![Histograms of Soil Pb concentrations](image)

**Figure 4.1:** Frequency histograms of soil Pb values in (A) Paisley, (B) Bishopbriggs, and (C) the Greater Glasgow G-BASE dataset. Note x axis values are matched but y axis varies between study areas due to different sample sizes.
Table 4.1: Summary statistics for soil Pb concentrations (mg kg⁻¹) in Greater Glasgow (G-BASE), Paisley, and Bishopbriggs. Comparison values are also provided.

<table>
<thead>
<tr>
<th>Soil Pb concentration (mg kg⁻¹)</th>
<th>Study area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Greater Glasgow</td>
</tr>
<tr>
<td>Mean</td>
<td>199</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>357</td>
</tr>
<tr>
<td>Median</td>
<td>135</td>
</tr>
<tr>
<td>Med.AD</td>
<td>78</td>
</tr>
<tr>
<td>Minimum - Maximum</td>
<td>17 - 9,937</td>
</tr>
<tr>
<td>Number of samples</td>
<td>1,285</td>
</tr>
</tbody>
</table>

World geochem. background: Mean = 25 (Matschullat et al., 2000)
Scotland mineral topsoil: Mean = 32; Median = 23 (Paterson et al., 2003)
England & Wales topsoil: Mean = 74; Median = 40 (McGrath & Loveland, 1992)

In the Glasgow G-BASE dataset soil Pb levels are elevated, as the median (135 mg kg⁻¹) is over five times greater than in Scottish rural soils (23 mg kg⁻¹; Table 4.1). Similarly, Fordyce et al. (2017) noted a three times enhancement in urban topsoil Pb levels in the Clyde Valley (which is a larger dataset including Greater Glasgow) compared with the average value for England and Wales. They attributed higher soil Pb levels in the Clyde Valley with anthropogenic pollution (Fordyce et al., 2017).

In Paisley and Bishopbriggs the topsoil Pb concentrations are also elevated, with the median concentrations (151 and 170 mg kg⁻¹) over seven times the Scottish rural median (23 mg kg⁻¹; Table 4.1). Soil Pb concentrations in Bishopbriggs have a larger span and median, but a lower mean, compared with Paisley (Table 4.1). In Bishopbriggs there are two extreme high measurements (Pb > 2,000 mg kg⁻¹; Figure 4.1B) whereas, in Paisley there are no such extremely high values, but there are more samples with relatively high Pb (i.e. Pb >1,000 mg kg⁻¹; Figure 4.1A).

Compared with Greater Glasgow, Bishopbriggs and Paisley have much higher mean soil Pb levels (>100 mg kg⁻¹ higher), but only slightly higher median Pb concentrations (16-35 mg kg⁻¹ increase), and a much narrower span (approximate fourfold lower maximum value, Glasgow = 9.937 mg kg⁻¹, Paisley = 1.766 mg kg⁻¹, Bishopbriggs = 2.422 mg kg⁻¹; Table 4.1). The large span in the G-BASE dataset is driven by two points with >5,000 mg kg⁻¹ Pb (Figure 4.1C). Which is to be expected as the G-BASE dataset covers a much wider area, encompassing a greater variety of land uses containing more extreme values. On the other hand, Paisley and Bishopbriggs are smaller datasets, hence soil Pb hotspots have a greater influence on mean values.
The high Med.AD (69-78 mg kg\(^{-1}\)) and standard deviation (357-383 mg kg\(^{-1}\)) in all locations indicates a high degree of variability in soil Pb levels. However, the Med.AD and standard deviation are similar in all areas (Table 4.1), indicating a comparable degree of variability about the median or mean, regardless of location.

### 4.3.2. Skew

Most statistical analyses require a Normal or Gaussian distribution of data; therefore, skew statistics and transform values were investigated in the city and neighbourhood datasets (Table 4.2).

<table>
<thead>
<tr>
<th>Table 4.2: Skew and lambda statistics for soil Pb datasets in Greater Glasgow (G-BASE), Paisley, and Bishopbriggs, both pre- and post-transform (Box-Cox).</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study Area</strong></td>
</tr>
<tr>
<td><strong>Pre</strong></td>
</tr>
<tr>
<td>Skewness</td>
</tr>
<tr>
<td>Octile skew</td>
</tr>
<tr>
<td>(\lambda) (optimal)</td>
</tr>
<tr>
<td><strong>Post</strong></td>
</tr>
<tr>
<td>Skewness</td>
</tr>
<tr>
<td>Octile skew</td>
</tr>
</tbody>
</table>

In the Glasgow G-BASE dataset a few outliers are causing a strong positive skew in the Pb data (Figure 4.1C), which is typical of a soil geochemical property (Lark and Lapworth, 2012). The presence of skew is reinforced by a high skewness value (18.33) and an octile skew >0.2 (0.46; Table 4.2). In Greater Glasgow the standard skew is much larger than octile skew. This indicates that it is mostly outliers in the upper octile that are leading to asymmetry in the distribution. In contrast, in the other study areas the octile and standard skews are more similar indicating that there are more numerous larger values leading to asymmetry in the distribution.

In Paisley and Bishopbriggs the octile skew is almost double that of the Glasgow G-BASE dataset (Table 4.2). However, these areas were chosen for their high range in soil Pb concentrations and are smaller datasets. There is minimal difference between octile skew values in Paisley and Bishopbriggs (Table 4.2), likely because they are similar sized datasets, both with some high outliers.

The optimal Box-Cox transformation lambda was determined for each study area using a restricted maximum likelihood (REML) model with constant fixed effects (lambda can also by fit based on covariates). After a Box-Cox transform the octile and traditional skew values are more acceptable i.e. ≤0.2 for octile skew for all locations apart from Paisley, and the traditional skew is close to 0 (Table 4.2).
4.3.3. Isotropy

To assess the stationarity and isotropy of the datasets the relationship between the Box-Cox transformed soil Pb value and Easting or Northing was plotted and modelled (Figure 4.2). Furthermore, multi-directional variograms were examined for each study area (Figure 4.3).

In the Glasgow G-BASE dataset there is no clear spatial trend in soil Pb concentrations, along either a Northing or Easting axis (Figure 4.2A). This lack of correlation is demonstrated by very low R² values and high p-values from the regression models (Figure 4.2A). However, the multi-directional plots indicate an unexpectedly high semivariance value for points approximately 100 m apart and at a 90° angle (Figure 4.3A). This is not reflected in any other direction. Nevertheless, this is a short lag distance so there may be a limited number of samples pairs this distance apart. Consequently, any outliers will skew the average semivariance. Generally, the variograms follow a similar pattern, so soil Pb in the Glasgow G-BASE dataset is mostly isotropic.

In Paisley and Bishopbriggs the soil Pb levels appear constant with direction, apart from a slight increase in soil Pb concentrations towards the southeast in Paisley (Figure 4.2B) and the north in Bishopbriggs (Figure 4.2C). When models are fitted to these relationships the correlation is extremely weak (R²<0.1), even when it is significant (p<0.05) (Figure 4.2B and C). It should be noted that these tests of the strength of relationship assume independent residuals whereas in this work spatially correlated residuals are expected and explored in the next chapter. Hence the significance of the correlatons with Northing or Easting are likely to be exaggerated.

Similar directional variograms support this conclusion (Figure 4.3). However, there is unusually high semivariance at low lag distances for 90° pairs in Paisley (Figure 4.3B), and in Bishopbriggs there is little semivariance at the first lag distance (35 m) for all directions apart from 0° (Figure 4.3C). Again, this may be due to fewer pairs of measurements at this short distance and direction. Overall, some degree of variation is expected in soil Pb levels, but a coordinate transform would only be required if a clearer relationship was present.
Figure 4.2: Soil Pb (Box-Cox) plotted against Eastings or Northings in (A) Greater Glasgow, (B) Paisley, and (C) Bishopbriggs. Linear regression models (fitted using OLS) are shown as red dashed lines, with their $R^2$ and $p$ values shown in red text above the plots. * Significant at >95% confidence, at **>99% confidence.
Figure 4.3: Multi-directional variograms for (A) Greater Glasgow, (B) Paisley, and (C) Bishopbriggs. Each colour represents average semivariance of pairs of sample points separated by a specific lag distance and direction.
4.4. Conclusions

In summary, the soil Pb concentrations are elevated above their expected background level in all datasets, due to enhanced anthropogenic pollution in urban environments. The specific cause of these elevated soil Pb concentrations are explored in Chapters 6 and 7. Soil Pb concentrations are slightly higher in the neighbourhood datasets, compared with the larger city-wide dataset. However, the Greater Glasgow dataset contains more extremely high outliers. All datasets are positively skewed, so it is recommended that they are transformed before any statistical analysis. A Box-Cox transformation generates a more Normal distribution. Lastly, the isotropy of the datasets was explored, and although direction might have a slight influence on soil Pb concentrations, it is not extreme enough to merit a coordinate transform or anisotropic variogram model.

References - Chapter 4


Appendix 4.1: Exploratory analysis R code

# N.B. "..." denotes where analysis only shown for one dataset but performed on all three datasets.

# Read in the data:
BB.df <- read.csv("BB_meanPb.csv", header=T)

# Pb summary statistics:
describe(BB.df$mean.Pb)
Octiles <- quantile(BB.df$mean.Pb, probs=c(0.125, 0.5, 0.875), Type=8)
Oc1 <- as.numeric(Octiles[1])
Oc4 <- as.numeric(Octiles[2])
Oc7 <- as.numeric(Octiles[3])
print(Oc1)

# Frequency histograms:
Hist(BB.df$Pb, breaks=200, main="", xlab="Pb (mg kg-1)", col="grey")

# Box-Cox transform:
bc <- boxcox(mean.Pb~1, data=BB.df, lambda=seq(-1.5,1.5,0.01))
print(lambda <- bc$x[which.max(bc$y)])
BB.df$bcPb <- ((BB.df$mean.Pb^lambda)-1)/lambda

# Spatial trend, using OLS, estimate the spatial trend in each direction:
trend.modelE.BB <- lm(BB.df$bcPb ~ BB.df$easting)
trend.modelN.BB <- lm(BB.df$bcPb ~ BB.df$northing)
summary(trend.modelE.BB)
summary(trend.modelN.BB)

# Plot northing/easting vs bcPb with regression lines:
par(mfrow=c(3,2))
plot(BB.df$northing, BB.df$bcPb, xlab="Northing", ylab="bcPb", pch=19)
abline(trend.modelN.BB, col="red", lty=2)
plot(BB.df$easting, BB.df$bcPb, xlab="Easting", ylab="bcPb", pch=19)
abline(trend.modelE.BB, col="red", lty=2)

# Multi-directional variograms. First make a geodata object:
BB.geo <- as.geodata (BB.df, coords.col=10:11, data.col=21)

# Plot a directional variogram:
emBB.var4 <- variog4 (BB.geo, estimator.type="modulus", uvec=10, max.dist=350)
plot(emBB.var4, xlab="Distance (m)", ylab="Semivariance")
Chapter 5. Comparing the spatial dependence and variability of soil Pb at different scales.

The Greater Glasgow G-BASE dataset and the two neighbourhood datasets (Paisley and Bishopbriggs) were used to examine the spatial dependence and variability of Pb in urban soils. This chapter examines the spatial dependence of soil Pb values (using Global and Local Moran’s I analyses), and the spatial variability of soil Pb concentrations (via semivariograms) at different scales. Furthermore, ordinary kriging was used to map spatial patterns in soil Pb concentrations. Understanding soil Pb’s spatial dependence, variability, and patterns is fundamental in order to identify controls or covariates associated with these patterns (Chapters 6 and 7). The use of more traditional geostatistical mapping methods i.e. kriging, is also required in order to compare the results with machine learning predictions (Chapter 8).

5.1. Background

Geochemical maps are fundamental for communicating the scale and amount of variation in a contaminant, such as Pb. The maps display soil Pb’s spatial variability, which can consequently influence sampling strategies, mapping accuracy, predictive modelling, risk assessment, and any future remediation project.

The most straightforward method for mapping soil Pb concentrations is to plot the data as discrete point maps using either graduated or varying colour or size of symbols (Longley et al., 2005). The benefit of these discreet-object maps is that they truly represent the element concentration measured at each sample point. However, it is often difficult to discern geochemical patterns and relationships on point maps. Therefore, it is common in geochemistry to convert point data to a continuous surface by modelling concentrations at unknown locations using interpolation from known data points.

There are many different types of interpolation techniques available. Thiessen polygons simply stratify space into individual polygons around a sample point, with the polygon given the class or value at the sample location (Oliver and Webster, 2015). This generates a continuous map but often with sharp boundaries between classes. Inverse distance weighting (IDW) is a more sophisticated method, as it draws on Tobler’s first law of geography - that objects closer in space are more
similar than those further apart (Tobler, 1970). Inverse distance weighting predicts a value at an unknown location by averaging nearby sample measurements, with more weight given to closer points (Longley et al., 2005). This generates a smoother map but can miss peaks and troughs in the data and does not provide uncertainty measurements (Longley et al., 2005).

Kriging is a common and robust method for interpolating and mapping a variable (Longley et al., 2005). There are many different types of kriging methods available, but all rely on the solid theoretical principles of geostatistics (Oliver and Webster, 2015). Kriging predictions are often more realistic than other interpolation methods and have the added advantage of measuring uncertainty alongside predicted values (O'Sullivan and Unwin, 2010).

Semivariograms are a fundamental part of kriging, as they model the variation in an element over different distances and at varying scales. Specifically, semivariograms quantify the relationship between sample separation distance (i.e. lag distance \( h \)) and semivariance (i.e. similarity between measurements \( \gamma(h) \)) in the concentration of an element (i.e. soil Pb \( z \)) at pairs of points \( (x_i, x_i + h) \); Diggle and Riberio, 2007):

\[
\gamma(h) = \frac{1}{2N(h)} \sum_{i=1}^{N(h)} [z(x_i) - z(x_i + h)]^2 ,
\]

(5.1)

where \( N(h) \) is the number of pairs of points separated by lag distance \( h \).

Semivariogram models are widely used to calculate the spatial range of soil Pb (i.e. the maximum separation distance over which pairs of measurements are correlated), and when approximated by a mathematical model they are used to spatially interpolate soil Pb concentrations over a given area. Many different methods, on multiple software platforms, are available for generating and fitting semivariogram models. Subsequently, semivariograms are often used to krige soil Pb values. However, subjective choices made by the user during the semivariogram generation and fitting procedures can control the appearance of the spatially interpolated maps (Marchant, 2018).

Traditionally, the modelling of spatial variability of an element, and prediction at unknown locations, would be achieved by classical geostatistics (Lark and Minasny, 2018). Classical geostatistics uses the method of moments (MOM) estimator (Oliver
and Webster, 2017) to average the semivariance for pairs of measurements separated by different lag distances. The user selects a series of lag bins which are defined according to the separation distance between pairs of measurement locations and, possibly, the direction between these locations. Tolerances are placed on these lag distances and directions, and the average semivariance for pairs of points falling into each bin is calculated.

This generates an empirical or experimental semivariogram which can be presented as plots of semivariance against lag distance in the directions considered. Different models are then fitted to this empirical semivariogram using a regression approach e.g. weighted least squares (WLS). These models represent the spatial variation in a variable using parameters such as those shown in Figure 5.1 (Oliver and Webster, 2015). Importantly, the sill \( (c + c_0) \) represents the total amount of variance in the dataset, the nugget \( (c_0) \) signifies measurement error or variation at a scale too small to be discernible from the sample design, the partial sill \( (c) \) indicates the amount of spatially correlated variance, the smoothness \( (v) \) shows how smoothly the variogram approaches the origin or nugget, and the range \( (a) \) is the lag distance at which the semivariogram reaches the sill, after this value the points are no longer spatially correlated (Diggle and Ribeiro, 2007; Marchant, 2018).

Different models e.g. spherical, exponential, or Gaussian, are often tested, with the optimum model used to krig the dataset. One particularly flexible model is the Matérn model, which is popular because it uses the smoothness parameter \( (v) \) to model varied semivariance values at short lags (Marchant and Lark, 2007). In some models the variogram increases asymptotically to the sill, so the effective range is used. The effective range is when the semivariance value is 95% of the sill (Refaat, 2007).

Although classical geostatistics is widely used to estimate soil Pb concentrations at unknown locations, Marchant (2018) highlights several issues associated with this
method. Firstly, choices made in the variogram estimation process can be subjective and susceptible to user error. For instance, the lag direction, distance, and maximum lag value, are important when generating the empirical semivariogram and are selected by the user. Similarly, the semivariogram model type, regression weighting function, and starting parameters (i.e. initial partial sill and range values) are key when model fitting but are not automatically determined. These often require some previous knowledge about the dataset, to prevent a ‘local minima’ being reached over the global ‘optimum’ value. Secondly, WLS relies on the assumptions that the data follows a Gaussian weakly stationary or isotropic pattern. Most soil Pb data does not meet the first criteria of Normal distribution, but this can be corrected for by transforming the dataset (Chapter 4). The second criteria of isotropy can be tested for and corrected by transforming the coordinates if required.

More recently, model-based geostatistics has become popular because it estimates a semivariogram model directly from the data, thereby omitting the subjective step of calculating an empirical semivariogram (Olea, 1994; Marchant 2018). This means the semivariogram model parameters fit the data instead of average bin values. Overall, there is less user control, although some of the properties can be pre-set. Generally, in model-based geostatistics a numerical optimization algorithm is used to find the semivariogram model parameters (and transformation parameters – see Equation 4.2) that maximise the likelihood that the observed data would have arisen from the fitted model (Marchant, 2018).

Previous urban studies indicate a large disparity in the spatial range of soil Pb i.e. the maximum distance at which two samples are spatially correlated (Table 5.1).
Table 5.1: Studies using variograms to measure soil Pb’s spatial variation, with study details, soil Pb summary statistics, kriging and validation methods.

Sorted by study date (see references at end of table). Abbreviations: NR = not reported, C.G. = classical geostatistics e.g. WLS. Spher. = spherical, Expo. = exponential, Gaus. = Gaussian model. OK = ordinary kriging, LogK = log-normal kriging, IK = indicator kriging, CS = conditional simulation (alternative to kriging). LOOCV = leave-one-out cross validation. \( \hat{\theta} \) median SSE, \( \bar{\theta} \) mean SSE (see Equation 5.10). RMSE = root mean square error (see Equation 8.3). (E-)BLUP = (empirical) best linear unbiased predictor (see Equation 6.5) \* Approximates as estimated from figure in paper.

<table>
<thead>
<tr>
<th>Study location</th>
<th>Samples</th>
<th>Soil Pb (mg kg(^{-1}))</th>
<th>Variogram parameters</th>
<th>Kriging</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scale (m)</td>
<td>N</td>
<td>Mean</td>
<td>Med.</td>
<td>Min. - Max.</td>
</tr>
<tr>
<td>Glebe, Australia(^1)</td>
<td>100</td>
<td>434</td>
<td>854</td>
<td>316</td>
<td>10 - 20,278</td>
</tr>
<tr>
<td>Silvermine, Ireland(^2)</td>
<td>400</td>
<td>218</td>
<td>NR</td>
<td>154</td>
<td>25 - 14,842</td>
</tr>
<tr>
<td>Sheffield, UK(^3)</td>
<td>500</td>
<td>569</td>
<td>244</td>
<td>164</td>
<td>19 - 4,300</td>
</tr>
<tr>
<td>Wolverhampton, UK(^4)</td>
<td>200-500</td>
<td>454</td>
<td>217</td>
<td>249</td>
<td>27 - 2,853</td>
</tr>
<tr>
<td>Paris, France(^5)</td>
<td>16,000</td>
<td>72</td>
<td>34</td>
<td>23</td>
<td>14 - 254</td>
</tr>
<tr>
<td>Changxin, China(^6)</td>
<td>4,000</td>
<td>665</td>
<td>33</td>
<td>NR</td>
<td>16 - 63</td>
</tr>
<tr>
<td>Swansea, UK(^7)</td>
<td>250</td>
<td>373</td>
<td>432</td>
<td>224</td>
<td>NR</td>
</tr>
</tbody>
</table>

\( \hat{\theta} \): \( \bar{\theta} \) median SSE, \( \bar{\theta} \) mean SSE (see Equation 5.10). RMSE = root mean square error (see Equation 8.3). (E-)BLUP = (empirical) best linear unbiased predictor (see Equation 6.5) \* Approximates as estimated from figure in paper.
<table>
<thead>
<tr>
<th>Study location</th>
<th>Samples</th>
<th>Soil Pb (mg kg⁻¹)</th>
<th>Variogram parameters</th>
<th>Kriging</th>
<th>Validation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Scale (m)</td>
<td>N</td>
<td>Mean</td>
<td>Med.</td>
<td>Range</td>
</tr>
<tr>
<td>Garden, Detroit, US⁵</td>
<td>1 - 10</td>
<td>80</td>
<td>151</td>
<td>116</td>
<td>17 - 882</td>
</tr>
<tr>
<td>Tehran, Iran⁶</td>
<td>250</td>
<td>116</td>
<td>21</td>
<td>NR</td>
<td>12 - 40</td>
</tr>
<tr>
<td>Pb-Zn mine, China¹⁰</td>
<td>Varied</td>
<td>97</td>
<td>54</td>
<td>286</td>
<td>12 - 37</td>
</tr>
<tr>
<td>Baia Mare, Romania¹¹</td>
<td>NR</td>
<td>120</td>
<td>864</td>
<td>298</td>
<td>38 - 14,329</td>
</tr>
<tr>
<td>Sydney, Australia¹²</td>
<td>Varied</td>
<td>1,122</td>
<td>577</td>
<td>161</td>
<td>3 - 20,278</td>
</tr>
<tr>
<td>Smelter, S. Korea¹³</td>
<td>5000</td>
<td>153</td>
<td>197</td>
<td>161</td>
<td>14 - 962</td>
</tr>
<tr>
<td>SW France¹⁴</td>
<td>16000</td>
<td>356</td>
<td>NR</td>
<td>27</td>
<td>3 - 261</td>
</tr>
</tbody>
</table>

References: ¹Markus & McBratney, 1996 & 2001; ²McGrath et al., 2004; ³Rawlins et al., 2005; ⁴Hooker and Nathanail, 2006; ⁵Saby et al., 2006; ⁶Shi et al., 2007; ⁷Marchant et al., 2011; ⁸Bugdalski et al., 2014; ⁹Mahmoudabadi et al., 2015; ¹⁰Yan et al., 2015; ¹¹Chakraborty et al., 2017; ¹²Johnson et al., 2017; ¹³Kim et al., 2019; ¹⁴Savignan et al., 2021.
Some studies have observed short range variation in soil Pb. For example, spatial variation in soil Pb was observed at the site scale, 7.5 m², and was attributed to land use history in Oakland, California (McClintock, 2012). Another study used variograms to show that soil Pb levels varied at the very fine scale (1 - 10 m) in one community garden in Detroit (Bugdalski et al., 2014). However, other studies have found that soil Pb has spatial pattern on a much larger scale/over several kilometres (McGrath et al., 2004; Rawlins et al., 2005; Savignan et al., 2021; Table 5.1).

This large disparity in the range of soil Pb variation is partly due to the lack of a standardised practise for analysing a dataset, i.e. different methods, platforms, and models are used in Table 5.1. However, the variation in soil Pb range is also dictated by the sampling interval, as studies with a larger sampling interval tend to find a larger spatial range in soil Pb (Table 5.1). This is likely because the resolution of their dataset is not fine enough to observe small scale variation in soil Pb. Some studies overcome this issue using multiscale or nested sampling (e.g. Hooker and Nathanail, 2006). The degree of Pb contamination and the presence of outliers are also important, despite most studies in Table 5.1 transforming their datasets.

Only a few studies have used model-based geostatistics to examine the variance in soil Pb (e.g. Marchant et al., 2011; Johnson et al., 2017). The lack of widescale use of model-based geostatistics to analysis soil Pb is sometimes due to dataset size requiring large computational power (e.g. Lark and Scheib, 2013). However, numerous studies have used model-based geostatistics to generate semivariograms for soil pH (e.g. Laslett and McBratney, 1990; Haskard et al., 2007; Makungwe et al., 2021) and other PHEs (e.g. arsenic (Niazi et al., 2011), cadmium (Marchant et al., 2011), and zinc (Minasny and McBratney, 2007)).

5.2. Chapter aims

Given the previous literature and the thesis goals outlined in Chapter 1, the aims of this chapter are to determine:

1. Does the spatial dependence and variability of soil Pb concentrations vary between the city and neighbourhood scale?

2. What is the most suitable method to map soil Pb at the city and neighbourhood scale?
5.3. **Methodology**

5.3.1. **Spatial dependence**

Tobler’s first law of geography states that objects close together in space will be more similar than those further apart (Tobler, 1970). A dataset is said to have positive spatial dependence, or autocorrelation, if measurements made close together in space are more similar than would be expected from an independent and random distribution of values (Haining et al., 2010). The presence of spatial dependence means that neighbouring measurements may not always provide substantial extra information and, if the study area is not sampled at random, the variance of the measured values can be poorly estimated (Haining et al., 2010). Therefore, a dataset’s spatial dependence should be assessed before any statistical analysis is performed.

Global Moran’s I (Moran, 1950) is a commonly used indicator to assess the degree of spatial autocorrelation or dependence in a dataset. Moran’s I operates by calculating the observed I value ($I$) for the dataset (O’Sullivan and Unwin, 2010). When calculating this value the spatial weights matrix ($W$) is fundamental (Getis, 2010). It stores information on the spatial relationship between neighbouring points and can vary considerably depending on the method chosen to assign neighbours (O’Sullivan and Unwin, 2010). In this study an inverse distance dependent matrix was used. This uses a distance decay function [$w_{ij} = d_{ij}^{-a}$] where the weight at $ij$ is the inverse distance between the points, weighted by a decay variable ($a$; Getis, 2010).

To calculate the observed I value ($I$) the soil Pb concentrations at points $i$ ($x_i$) and $j$ ($x_j$) are compared with the global average of all points ($\bar{x}$; O’Sullivan and Unwin, 2010). This covariance is multiplied by a component from the spatial weights matrix ($w_{ij}$) and divided by the total spatial weights of the dataset ($S_0$; O’Sullivan and Unwin, 2010). Lastly, the resulting value is normalised by the overall variance of the dataset ($n/\sum(x_i - \bar{x})^2$):

$$I = \left[ \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{S_0} \right] \left[ \frac{n}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \right].$$

(5.2)
If the I value is >0 then there is positive spatial dependence in the dataset, with any I values >0.3 showing strong spatial dependence (O’Sullivan and Unwin, 2010).

The expected I value \( E(I) \) is used to calculate the \( p \)-value and z-score associated with the observed I value i.e. the statistical significance of any spatial dependence. It is dependent on the number of sample points \( n \); O’Sullivan and Unwin, 2010:

\[
E(I) = \frac{-1}{n - 1}.
\]  

(5.3)

Global indicators of spatial dependence assume the spatial dependence is homogenous across the whole dataset, whereas local indicators assess the heterogeneity of spatial dependence (Getis, 2010). Specifically, Local Indicators of Spatial Autocorrelation (LISA; Anselin, 1995) look at the spatial dependence between neighbouring samples or units. In this study LISA was used to calculate and map where soil Pb levels in neighbouring samples are spatially dependent.

LISA breaks down the global Moran’s I calculation into its local components, so it is useful for identifying spatial clusters in the dataset (Getis, 2010). The local Moran’s I value \( l_i \) is calculated in a similar way to the global Moran’s I \( I \) (Equation 5.2) but lacks the terms that normalise the I value for the whole dataset (Getis, 2010):

\[
l_i = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (x_i - \bar{x}) (x_j - \bar{x}) .
\]  

(5.4)

The associated \( p \)-value can then be calculated using random permutations, similar to Equation 5.3 (Getis, 2010):

\[
E(l_i) = -\frac{1}{n} \sum_{j=1}^{n} W_{ij} .
\]  

(5.5)

Global Moran’s I and LISA were performed on the Glasgow G-BASE dataset and the two neighbourhood datasets using the spdep package (Version 1.2.2) in R (Version 4.1.2), with example R script provided in Appendix 5.1. The global Moran’s I value assessed total spatial dependence in each dataset, with correlograms exploring the change in spatial dependence with increasing lag distances (Getis, 2010). The range in local Moran’s I value were examined then plotted in a scatterplot. A Moran’s I scatterplot compares the standardised soil Pb value, and
spatially lagged soil Pb value i.e. standardised soil Pb multiplied by its spatial weight ($W$) for each point (O’Sullivan and Unwin, 2010). Therefore, points with a high soil Pb value but also a high spatially lagged soil Pb value (Quadrant 1), have been influenced by neighbouring points with high soil Pb values, and vice versa for samples in quadrant 2. In quadrant 3, the soil Pb value at the measured sample is higher than average, but the surrounding neighbours have lower values, with the opposite for quadrant 4. This information was also mapped to identify clusters of points where the sampled concentrations of Pb exhibited positive spatial dependence.

5.3.2. Spatial variability

The spatial variability in each dataset was explored using omnidirectional semivariograms. Generally, there are two key options available for generating a semivariogram model and subsequently kriging a variable. Firstly, classical geostatistical methods or the WLS method, generates a semivariogram for several lags then estimates model parameters by minimising the squared difference between the semivariogram points and the model estimates (Jian et al., 1996). This can be expanded by using robust geostatistical methods, which minimise the effect of outlying measurements on the general trend when fitting a semivariogram. However, this is more applicable to rural regions where studies focus on mapping the underlying variation in soil PHEs, such as Pb. In urban studies, sites with large soil Pb concentrations are of interest, as they pose greater potential risk. Alternatively, model-based geostatistics or maximum likelihood methods uses the raw data directly to find the optimum model that maximises the likelihood function (Jian et al., 1996, Marchant, 2018). However, the data was transformed during model optimisation to ensure outlying values are consistent with the model.

Classical geostatistical methods using R (Version 4.1.2) and ArcGIS (Version 10.8.2, Esri®) were compared with model-based semivariogram fitting using the British Geological Survey’s Geostatistical Toolbox for Earth Scientists (Marchant, 2018) which is currently only available for Matlab (R2020a, MathWorks Inc.®). The results from this comparative study can be found in Appendix 5.2 (Table 5.7). In summary, cross-validation results indicate that model-based geostatistics generate the ‘best’
semivariogram models i.e. semivariogram models which produce the most accurate kriged predictions and measures of uncertainty.

To generate semivariogram models using model-based geostatistics the methodology outlined in Marchant (2018) was adapted (Matlab example code in Appendix 5.3). Firstly, a linear mixed model (LMM) predicted soil Pb concentrations at location \( x_i \) using \( s_i = s(x_i) \) and the set of observations \( s = \{ s_1, s_2, \ldots, s_n \}^T \) (\( T = \) the transpose of the vector). The LMM builds on the equation for a straight line (i.e. \( y = mx + c \)), by predicting \( s \) using:

\[
s = M\beta + \epsilon. \tag{5.6}
\]

Where \( M\beta \) are the fixed effects including \( \beta \) the mean soil Pb concentration for each sample \( (\beta_1, \beta_2, \ldots, \beta_q)^T \), and \( M \) the design matrix \( (\text{size} = n \times q, \text{when} q = \text{number of covariates}) \). Covariate values can be included in columns of the design matrix (see Chapter 6). The random effects (i.e. noise or intercept) are represented by \( \epsilon \).

This LMM was expanded to fit the variogram using REML (Patterson and Thompson, 1971). The Nelder and Mead (1965) method was run, which uses ten different starting points to avoid finding a local maximum (Marchant, 2018). Additionally, the site coordinates were rescaled to simplify computations and only unique locations were input. Restricted maximum likelihood finds the transformation and semivariogram parameter values (which are grouped together in vector \( \alpha \) ) values which maximise the log of the likelihood that the observed data would have arisen from the proposed model. The likelihood \( (L) \) is written (Marchant, 2018):

\[
L(s|\alpha) = \text{Constant} + \frac{1}{2} \ln|W| + \frac{1}{2} \ln|C(\alpha)| + \frac{1}{2} s^T C(\alpha)^{-1} Q s, \tag{5.7}
\]

where \( W = M^T C(\alpha)^{-1} M \) and \( Q = I - MW^{-1}M^T C(\alpha)^{-1} \).

In the above calculations, \( M^T \) is a transpose of the fixed effects design matrix \( M \) (see Equation 5.6), \( C(\alpha)^{-1} \) is the covariance matrix of random effects, and \( s \) is a transformation function (i.e. the Box-Cox function Equation 4.2; Marchant, 2018). Subsequently, the corresponding \( \beta \) value which maximises the likelihood function (Equation 5.7) can then be calculated from (Marchant, 2018):

\[
\beta = (M^T C(\alpha)^{-1} M)^{-1} M^T C(\alpha)^{-1} s. \tag{5.8}
\]
Where $\beta$ is the mean soil Pb concentration (as in Equation 5.6). In the implementation of the software used in this thesis, $\alpha$ contains six key semivariogram parameters, the nugget, partial sill, range, smoothness (Matérn), lambda, and measurement error (Marchant, 2018). Any of these values can be fixed. Due to previous issues with measurement error (see Appendix 5.2), this parameter was pre-set to 0. The design matrix $M$ consisted of a single column of ones so that the fixed effects were a constant. The other parameter values were optimised during modelling.

The optimised variogram parameters were output for each model, and the variogram was plotted. Additionally, MOM points i.e. average semivariance for specified bin widths, were added to the plot in order to assess the fit of the variogram with the actual data.

5.3.3. Mapping soil Pb

Kriging is often referred to as the empirical best linear unbiased predictor (E-BLUP) because it minimises the expected error upon prediction at an unobserved location (Oliver and Webster, 2015). There are different types of kriging, but ordinary kriging (OK) is arguably the most robust and commonly used type of kriging (Table 5.1). It predicts a parameter at an unknown location $[\hat{Z}(x_0)]$ using known data values $[z(x_i)]$, and kriging weights ($\omega$) generated by the semivariogram (Oliver and Webster, 2015):

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \omega_i z(x_i).$$

(5.9)

The manner in which the kriging weights are determined ensures that known data values in the neighbourhood of the unknown location have a greater influence on the predicted value (Haining et al., 2010). The data is assumed to be weakly stationary and isotropic. Otherwise, universal kriging can correct for a strong spatial trend (Lark and Lapworth, 2012).

In this study, OK predicted soil Pb concentrations for a 5 m by 5 m prediction grid in Paisley and Bishopbriggs, and a 200 m by 200 m prediction grid covering the Greater Glasgow study area (all grids >7000 points). Semivariogram parameters were used to krig soil Pb values, along with known sample concentrations.
Furthermore, 1000 kriging simulations were run, which generated kriging predictions in untransformed units and standardised error (see Appendix 5.2. Section 4 for details).

After kriging has been used to produce an interpolated surface, the semivariogram model and kriging accuracy must be considered. Ten-fold cross-validation was used to quantitively assess kriging accuracy. This was chosen over leave-one-out cross-validation (LOOCV) to reduce processing time. Ten-fold cross validation operates by randomly assigning each measurement to one of ten folds, the model is then re-run without this fold and the OK predicted and actual results are compared for fold values. This is repeated for each fold and the root mean squared error (RMSE) and bias are calculated. Another indicator of kriging fit is the mean and median squared standardised prediction error (SSE or $\bar{\theta}$ and $\tilde{\theta}$; Marchant, 2018). These are determined by squaring the difference between the actual value ($s_i$) and the predicted value ($\hat{s}_i$) for each point and dividing by the prediction variance ($V_i$) (Marchant, 2018):

$$\theta_i = \frac{(s_i - \hat{s}_i)^2}{V_i}.$$  \hspace{1cm} (5.10)

Accurate kriging predictions will have $\bar{\theta}$ close to 1 and $\tilde{\theta}$ close to 0.455 (Minasny and McBratney, 2007). As a further check, the histogram of the standardised squared errors was investigated to confirm they follow a standardised Normal distribution (i.e. mean=0 and variance=1), which also indicates a well fit model.

In order to best convey the resulting kriged values and their associated uncertainty, different communication methods were explored. The aim was to practically communicate the soil Pb variability to decision makers, while retaining uncertainty information. This included looking at soil guideline values (Section 2.3.), side-by-side displays, and transparency (MacEachren et al., 2005). All colour schemes followed ColorBrewer (Brewer, 2006) guidelines to help the user differentiate between different classes. The first maps (Figure 5.6) were created using ArcMap (10.8.2; Esri®). However, some display options i.e. varying layer transparency by an attribute, were only available in ArcGIS Pro (2.8.0; Esri®). Consequently, ArcGIS Pro was used to create the later maps (Figures 5.7 and 5.8).
5.4. Results

5.4.1. Spatial dependence

The spatial dependence in each dataset was quantified using global and local Moran’s I analysis. Before Moran’s I analysis the spatial weights matrix (W) was checked to ensure all samples had at least one neighbouring point. Table 5.2 shows this was the case (i.e. minimum neighbours = 1). Additionally, samples in the Greater Glasgow dataset tend to have more neighbours (mean = 7) separated by a larger distance (mean = 356 m), compared with Paisley and Bishopbriggs which are relatively similar (mean = 5 neighbours 28-29 m apart; Table 5.3). This reflects the different sampling scales for each study area (Chapter 3).

<table>
<thead>
<tr>
<th>Table 5.2: Summary statistics for the number of neighbours and the distance (m) between them in the spatial weights matrix for each study area.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NEIGHBOURS</strong></td>
</tr>
<tr>
<td>Min. – max.</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Mean</td>
</tr>
<tr>
<td><strong>DISTANCE (m)</strong></td>
</tr>
<tr>
<td>Min. – max.</td>
</tr>
<tr>
<td>Median</td>
</tr>
<tr>
<td>Mean</td>
</tr>
</tbody>
</table>

The global Moran’s I value determines the degree of spatial dependence in the whole dataset. In Greater Glasgow the resulting Moran’s I value was 0.15, which is significantly greater (p = 2.2e-16) than the expected Moran’s I value (Table 5.3) under a null model of independent measured values. This indicates there is positive spatial dependence in the dataset, however because 0.15 < 0.3 the dependence is classified as weak. Furthermore, the correlogram (Figure 5.2A) indicates that the degree of spatial dependence decreases with increasing lag distance, with weak positive spatial dependence occurring between points separated by up to 1.6 km. Points separated by 1.6 to 3.9 km have very weak spatial dependence, and after 3.9 km there is virtually no relationship between neighbouring points (Figure 5.2A).

<table>
<thead>
<tr>
<th>Table 5.3: Observed and expected Global Moran’s I values with their significance (p) values for each study area.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Greater Glasgow</strong></td>
</tr>
<tr>
<td>Moran’s I</td>
</tr>
<tr>
<td>Expected I</td>
</tr>
<tr>
<td>p</td>
</tr>
<tr>
<td>+0.15</td>
</tr>
<tr>
<td>0.00078±0.00027</td>
</tr>
<tr>
<td>&lt;&lt;0.01** (2.2e-16)</td>
</tr>
<tr>
<td><strong>Paisley</strong></td>
</tr>
<tr>
<td>Moran’s I</td>
</tr>
<tr>
<td>Expected I</td>
</tr>
<tr>
<td>p</td>
</tr>
<tr>
<td>+0.33</td>
</tr>
<tr>
<td>-0.009±0.0049</td>
</tr>
<tr>
<td>&lt;&lt;0.01** (1.1e-4)</td>
</tr>
<tr>
<td><strong>Bishopbriggs</strong></td>
</tr>
<tr>
<td>Moran’s I</td>
</tr>
<tr>
<td>Expected I</td>
</tr>
<tr>
<td>p</td>
</tr>
<tr>
<td>+0.21</td>
</tr>
<tr>
<td>-0.0073±0.00032</td>
</tr>
<tr>
<td>&lt;&lt;0.01** (1.2e-4)</td>
</tr>
</tbody>
</table>
The global Moran’s I value also reveals positive spatial dependence in the neighbourhood datasets (Table 5.3). In Paisley the Moran’s I value is >0.3 (+0.33) and significant ($p$<<0.01), indicating strong positive correlation. Although, the correlogram suggests this only occurs for samples <54 m apart (Figure 5.2B). In Bishopbriggs the Moran’s I value is significant ($p$<0.01), but it is lower than in Paisley and <0.3 (+0.21), implying weaker spatial dependence. As with Paisley, this spatial dependence is only present as short lag distances (<59 m; Figure 5.2C).
Figure 5.2: Correlograms showing Moran’s I value for different lags distances (m). Black dot equals mean, with upper and lower confidence intervals (95%). Lag distance equals maximum neighbour separation distance (Table 5.2).
Local Moran’s I analysis shows that the degree of spatial dependence varies spatially throughout each dataset. In Greater Glasgow LISA ranges from -3.26 to +4.9, with a quarter of samples having strong positive spatial dependence (i.e. I > +0.3; Table 5.4). In the neighbourhood datasets the local observed Moran’s I ranges from -2 to +2.77, with almost half (46%) of samples in Paisley showing strong positive spatial dependence (Table 5.4). However, the spatial correlation in these points is not always significant. This is demonstrated by the wide range in p values (0 - 1) in each dataset, with only 8-14% of samples having significant dependence (i.e. p < 0.05; Table 5.4).

| Table 5.4: Summary of LISA statistics for observed Moran’s I value (I) and p value for each dataset (the mean LISA value equates to the global Moran’s I value; Table 5.3). The number of samples with strong positive (I; > +0.3) and significant (p < 0.05) spatial dependence are also given. |
|---------------------------------|-----------------|-----------------|-----------------|
| **Observed I**                  | Greater Glasgow | Paisley         | Bishopbriggs    |
| Min. - max.                     | -3.26 to +4.9   | -1.01 to +2.09  | -2.02 to +2.77  |
| Median (mean)                   | +0.05 (+0.15)   | +0.22 (+0.33)   | +0.02 (+0.21)   |
| **p value**                     |                 |                 |                 |
| Min. - max.                     | 0 - 0.997       | 0.011 - 0.999   | 0 - 0.996       |
| Median (mean)                   | 0.37 (0.41)     | 0.3 (0.38)      | 0.47 (0.46)     |
| **Number of samples (% total) with spatial dependence that is:** |                 |                 |                 |
| Strong positive                 | 320 (25%)       | 52 (46%)        | 36 (26%)        |
| Significant                     | 160 (12%)       | 9 (8%)          | 19 (14%)        |
| Strong and signif.              | 95 (7%)         | 5 (4%)          | 13 (9%)         |

The span in LISA values were explored using Moran’s scatterplots (Figure 5.3) and maps (Figure 5.4). The Moran’s table and scatterplots (Table 5.4 and Figure 5.3) show a relatively wide distribution of results for each study area, indicating a large range of spatial dependence for different soil Pb concentrations. Furthermore, few points have significant spatial dependence, matching the results in Table 5.4. The maps identify significant hotspots (red points) and coldspots (blue points) of positively autocorrelated soil Pb values (Figure 5.4). In Greater Glasgow, the hotspots and coldspots tend to be spatially clustered (Figure 5.4A). Conversely, points with negative spatial dependence, i.e. significant high-low or low-high points found in quadrants 3 and 4 (Figure 5.3), are less common and tend to be scattered (Figure 5.4A). Clusters of samples with significant positive spatial dependence can also be witnessed in Paisley (Figure 5.4B) and Bishopbriggs (Figure 5.4C). In Paisley there are no coldspots and few spatial outliers, which was also seen on the scatterplot (Figure 5.3B).
Figure 5.3: Moran’s scatterplots for standardised soil Pb (Box-Cox) versus spatially lagged soil Pb (Box-Cox i.e. standardised Pb * W) for each study area. Points with crosses have \( p<0.05 \). Quadrants match categories in maps (Figure 5.4). Quadrant 1, HH = high-high i.e. higher than average soil Pb and high surrounding values; quadrant 2, LL = low-low, i.e. lower than average soil Pb and low surrounding points; quadrant 3, HL = high-low i.e. higher soil Pb but low surrounding points; and quadrant 4, LH = low-high i.e. lower than average soil Pb with higher surrounding values. The best fit line is the red dashed line, with its equation shown in the bottom right and its slope equaling the Global Moran’s I value.
Figure 5.4: Maps of sample points in each study area coloured by Moran’s scatterplots quadrant where spatial dependence is significant (p<0.05; Figure 5.3). See Figure 5.3 caption for details on quadrants.
5.4.2. Spatial variability

The spatial variability in each dataset was explored using semivariograms (Table 5.5; Figure 5.5). In the Greater Glasgow dataset the semivariogram model has a much higher nugget than partial sill variance, as 75% of the total variance is composed of the nugget (Table 5.5). This makes the variogram a flatter and almost ‘pure nugget’ (Figure 5.5A). This is despite some MOM points at short lag distances (< 150 m) showing low variance, or high spatial dependence, that the model does not capture (Figure 5.5A). Other lag distances (i.e. 118 m and 180 m) with higher semivariance (i.e. > 0.06) may be skewing the average values. Overall, semivariogram models with a high nugget often arise from large amounts of measurement error due to a coarse sampling strategy.

In Paisley the MOM points indicate strong spatial dependence at the first lag distance (10 m), but increasingly less spatial dependence between points after 30 m, with an almost oscillating pattern (Figure 5.5B). Furthermore, the semivariogram model largely agrees with the MOM points. The Paisley model has a very low nugget to variance ratio (1%), as the nugget is very close to 0, with a significantly higher partial sill that reaches most the MOM points (Figure 5.5B; Table 5.5). This type of semivariogram model indicates that the sampling strategy sufficiently captured the underlying spatial variation.

In Bishopbriggs, the MOM points show low variance at the first lag distance (10 m), but this is not represented in the semivariogram model (Figure 5.5C). In the model the nugget is higher than the partial sill and makes up 59% of total variance (Table 5.5). This is likely due to two extremely high outliers separated by a short distance (~10 m). In Bishopbriggs the total semivariance is much lower than in Paisley (sill = 10x less). This is partially due to different transforms, i.e. lower lambda value in Bishopbriggs (Table 5.5). However, it may also arise because there is less spatial dependence in Bishopbriggs (Table 5.3).

| Table 5.5: Semivariogram model parameters for soil Pb in each study area. |
|-----------------------------|--------|--------|--------|
|                             | Greater Glasgow | Paisley | Bishopbriggs |
| Nugget                     | 0.04    | 0.00048 | 0.0023   |
| Partial sill               | 0.013   | 0.038  | 0.0016   |
| Sill                       | 0.053   | 0.04   | 0.0039   |
| Nugget/Sill %              | 75%     | 1%     | 59%      |
| Eff. Range (m)             | 10,506  | 71.76  | 115.82   |
| Smoothness                 | 0.28    | 1.5    | 1.5      |
| Lambda                     | -0.22   | -0.29  | -0.46    |
Figure 5.5: REML semivariograms modelling variance in soil Pb (Box-Cox) with lag distance. Green crosses are MOM points, showing average variance for pairs of points within each lag distance bin. In all locations the 1st bin centre is at 10 m, 2nd at 30 m, then at 30 m intervals.
5.3.3. Mapping soil Pb

Kriged soil Pb concentrations, and error, for Greater Glasgow, Paisley, and Bishopbriggs were displayed using three different mapping techniques.

Firstly, soil Pb predicted values in Box-Cox transformed units were mapped, with kriging error shown on a separate map (Figure 5.6). This allows the user to compare how the predicted values, and their uncertainties, vary between grid points (MacEachren et al., 2005). In the transformed unit maps, soil Pb break points for mapping classes were eight quantile breaks for each dataset. Consequently, mapping classes were not standardised across the three study areas. This was decided because of the large disparity in the range of kriged soil Pb values between areas (Figure 5.6), caused by differing lambda values (Table 5.5).

Secondly, kriged soil Pb concentrations in back-transformed units (i.e. mg kg\(^{-1}\)) were mapped (Figure 5.7). In these maps soil Pb classes were eight quantile breaks, but for a combined dataset. Consequently, class break points were standardised across all study areas, which provides greater comparability while still displaying the pattern in soil Pb concentrations. In the back-transformed unit maps prediction uncertainty was conveyed using transparency. Specifically, the transparency of each grid point was adjusted by kriging error (Box-Cox transformed so as not to disadvantage higher soil Pb values). Consequently, more uncertain predictions are more transparent, and vice versa. Furthermore, a darker basemap was selected for Figures 5.7 and 5.8 to emphasise the transparency of different point symbols.

Lastly, soil Pb concentration (mg kg\(^{-1}\)) were classified by the C4SL (see Section 2.3. for details) exceeded at each grid point. The C4SLs consider land use when determining potentially harmful soil Pb levels, with soil Pb trigger concentrations dependent on land use. In the kriged maps (Figure 5.8) the grid point is classified by land use at which the highest trigger value is exceeded by the predicted mean. However, this can mean multiple land use C4SLs are exceeded. For example, if a grid point is predicted to contain 800 mg kg\(^{-1}\) of soil Pb it will be classified as residential open space, but also exceeds the C4SL for allotments and gardens. As with the transformed unit maps, prediction uncertainty is communicated using transparency.
Figure 5.6A: On the left a map of kriged soil Pb concentrations (Box-Cox) for a 200 m x 200 m grid covering the Greater Glasgow study area. A map of kriging error (Box-Cox) is shown on the right.
Figure 5.6B: Upper map shows kriged soil Pb concentrations (Box-Cox) for a 5 m x 5 m grid covering the Paisley study area, with kriging error (Box-Cox) below.
Figure 5.6C: Upper map shows kriged soil Pb concentrations (Box-Cox) for a 5 m x 5 m grid covering the Bishopbriggs study area, with kriging error (Box-Cox) below.
Figure 5.7A: Map showing kriged soil Pb concentrations (mg kg\(^{-1}\)) for a prediction grid covering Greater Glasgow. Note same mapping classes used in all maps in Figure 5.7. Grid points with a higher kriging error (Box-Cox) have increased transparency.
Figure 5.7B: Map showing kriged soil Pb concentrations (mg kg\(^{-1}\)) for a prediction grid covering Paisley. Note same mapping classes used in all maps in Figure 5.7. Grid points with a higher kriging error (Box-Cox) have increased transparency.
Figure 5.7C: Map showing kriged soil Pb concentrations (mg kg\(^{-1}\)) for a prediction grid covering Bishopbriggs. Note same mapping classes used in all maps in Figure 5.7. Grid points with a higher kriging error (Box-Cox) have increased transparency.
Figure 5.8A: Map showing kriged soil Pb concentrations (mg kg$^{-1}$) for a prediction grid covering Greater Glasgow, classified by which land use(s) C4SLs are exceeded.

- None: <80 mg kg$^{-1}$,
- Allotment: 80 - 310 mg kg$^{-1}$,
- Garden: 310 - 630 mg kg$^{-1}$,
- Residential open space: 630 - 1300 mg kg$^{-1}$,
- Public open space: 1300 - 2300 mg kg$^{-1}$.

Grid points with a higher kriging error (Box-Cox) have increased transparency, hence appear darker.
Figure 5.8B: Map showing kriged soil Pb concentrations (mg kg\(^{-1}\)) for a prediction grid covering Paisley, classified by which land use(s) C4SLs are exceeded. See Figure 5.8A for class details. Grid points with a higher kriging error (Box-Cox) have increased transparency, hence appear darker.
**Figure 5.8C**: Map showing kriged soil Pb concentrations (mg kg$^{-1}$) for a prediction grid covering Bishopbriggs, classified by which land use(s) C4SLs are exceeded. See Figure 5.8A for class details. Grid points with a higher kriging error (Box-Cox) have increased transparency, hence appear darker.
5.3.4. Accuracy

In all study areas overall model validation statistics are reasonable, as the mean SSE is close to 1 (0 - 0.19 difference), and the median is not far from 0.455 (0.005 - 0.085 difference; Table 5.6). Paisley kriging estimates have exceptionally good cross validation statistics (i.e. mean SSE = 1, and median SSE very close to 0.45; Table 5.6), perhaps linked to the ideal semivariogram structure (i.e. low nugget). Greater Glasgow has good results, with the lowest RMSE, possibly due to more sample points. However, Bishopbriggs shows the worse validation statistics for all study areas, with mean SSE furthest from 1 and median SSE furthest from 0.455. This may be due to the influence of outliers.

Comparing cross-validation results with other studies is difficult due to differences in modelling techniques and transformations. Furthermore, there is no standardised validation method or output (Table 5.1). However, the results from this study are an improvement on Saby et al. (2006), who produced a median SSE of 0.34 and mean SSE of 0.83 but mapped soil Pb on a much larger scale (i.e. all of Paris). Lark and Scheib (2013) generated median SSEs ranging from 0.36 to 0.47 for different OLS semivariogram models of soil Pb in London. This is comparable with the range of median SSEs in this study.

| Table 5.6: Ten-fold cross validation statistics for REML fit variogram models with OK predictions for each study area. SSE = standardised squared error (see Equation 5.10; ŝ = mean, ideally close to 1; ŝ̅ = median, ideally close to 0.455). RMSE = root mean square error (in untransformed units; see Equation 8.3) |
|---|---|---|
| Greater Glasgow | Paisley | Bishopbriggs |
| Mean SSE (ŝ) | 1.03 | 1.00 | 1.19 |
| Median SSE (ŝ̅) | 0.37 | 0.46 | 0.49 |
| RMSE (mg kg⁻¹) | 115.87 | 193.13 | 221.17 |
5.5 Discussion

5.5.1. Spatial dependence

The semivariograms and Moran’s I value indicate spatial dependence is present in soil Pb concentrations in all three study areas. However, the amount and range of spatial dependence varies between locations. The spatial dependence is strongest in Paisley, as proven by its high and significant Moran’s I value ($I = +0.33$, $p << 0.01$; Table 5.3). This strong positive spatial dependence is not homogenous across the dataset, as there are localised hotspots where similar high soil Pb concentrations group together (Figure 5.4B). These hotspots could be attributed to point Pb pollution sources, perhaps from locations of former industry (see Chapter 6). The Moran’s I value in Paisley is similar to the Moran’s I value determined by Yan et al. (2015) of +0.38 for hotspots found around a Pb-Zn mine in southeast China. They also mapped LISA values to observe hotspots and coldspots of soil Pb concentrations, depending on proximity to the mine (Yan et al., 2015). Similarly, Zhao et al. (2019) determined a global Moran’s I value of +0.3 and identified hotspots and coldspots in their study of Pb concentrations in paddy soils around an electrical-waste site in southeast China. These two studies focused on sampled point locations in areas known to contain elevated levels of Pb pollution.

Soil Pb concentrations in Bishopbriggs and Greater Glasgow also show positive spatial dependence, but this is weaker than in Paisley. As indicated by the slightly lower, but still significant, Moran’s I values of +0.21 and +0.15 for Bishopbriggs and Greater Glasgow, respectively (Table 5.3). The Greater Glasgow dataset has the lowest $p$ value of any study area (Table 5.3), possibly due to larger sample numbers. Compared with Paisley, LISA maps in Bishopbriggs and Greater Glasgow contain more coldspots (Figures 5.4A and C). They also include more points with negative spatial dependence i.e. high-low or low-high points in quadrants 3 and 4 (Figures 5.3, A and C). This may indicate more spatial variability in these study areas.

Overall, spatial dependence is not greatly influenced by scale, as global and local Moran’s I statistics indicate that Bishopbriggs is more similar to Greater Glasgow, than to Paisley. Although more study areas would need to be investigated before a solid conclusion could be reached. Spatial dependence appears to be more influenced by number of samples (i.e. more samples causes more certain results),
and soil Pb's underlying spatial pattern (i.e. the presence of point Pb sources causing localised soil Pb hotspots).

5.5.2. Spatial range

Conversely, the range of spatial dependence, in other words soil Pb’s spatial variability, does vary with scale. This is demonstrated by comparing the correlograms (Figure 5.2) and their inverse, the semivariograms (Figure 5.5), for the neighbourhood and city datasets.

The correlograms indicate high spatial dependence in neighbouring points separated by shorter lag distances in all datasets (Figure 5.2). However, in the neighbourhood datasets the lag distance is 54 - 59 m, with neighbouring samples showing strong positive spatial dependence at the first lag distance only (Figures 5.2B and C). In the city dataset the lag distance is \(~10\)x greater (784 m), and neighbouring samples have positive spatial dependence up to 1.6 km apart (Figure 5.2A). The lag distance is the maximum neighbour sample distance (Table 5.2), so depends on the spatial weights matrix and the sampling strategy. Consequently, samples closer together than 784 m in the Greater Glasgow dataset may have strong spatial dependence, but there is not enough data to reach this conclusion.

Similarly, in the semivariogram models the effective range quantifies the distance up to which samples are spatially dependent. In the neighbourhood datasets the semivariogram models’ effective range is 72 - 116 m, whereas it is approximately 100x greater in the city dataset (10,506 m; Table 5.5). Oliver and Webster (2015) suggest the sampling interval should be approximately half the effective range in order to capture the correlated structure of a variable. In Paisley and Bishopbriggs the average sampling separation distance was 29 m (Table 5.2), so the semivariograms portray the spatial variation in soil Pb concentrations. Overall, in the neighbourhood datasets sampling separation should be on the scale of <60 m if soil Pb variation is to be sufficiently explored and mapped.

In the city dataset the average sample distance (356 m) is less than half the range (10,506 m), so the spatial variation in soil Pb should be modelled according to Oliver and Webster’s rule. However, the semivariogram has a very high nugget (75% of sill) indicating Pb spatial variability has not been fully captured. Other studies with similar sampling intervals observed a comparable range e.g. a range
of 3.9 km from a 500 m sampling interval (Rawlins et al., 2005) and 4.3 km range with a sampling distance of 400 m (McGarth et al., 2004). Although it is difficult to directly compare studies due to different variogram fitting methodologies.

5.5.3. Mapping soil Pb’s spatial variability

5.5.3.1. Classification breaks and units

Reimann (2005) highlights the importance of selecting appropriate mapping representations to convey spatial structures in an element and aid interpretability. Specifically, he warns that a suitable choice of colour, symbol size and type, classification breaks and units, should be made. He suggests that quantile-based intervals are best suited to mapping geochemical data.

This was explored in the Greater Glasgow and Paisley/Bishopbriggs datasets. Despite differences in classification breaks and units, the soil Pb kriged maps (Figures 5.6 - 5.8) show the same pattern in each study area. For example hotspots corresponds with hotspots and low concentrations with other low concentrations. This is unsurprising as all maps use the same dataset.

Differences in classification breaks can emphasise areas in one map and make them appears less important in another. For example, the highest soil Pb values in Bishopbriggs occur in Etive Crescent Park, in the northeast of the study area. In the Box-Cox transformed (Figure 5.6C) and back-transformed (Figure 5.7C) unit maps the hotspot at Etive Crescent Park is evident but is somewhat diluted by several other areas of concern. Conversely, in the C4SL map (Figure 5.8C) the hotspot at Etive Crescent Park is immediately detectable as the most concerning location.

Choice of classification breaks can also homogenise the underlying pattern in soil Pb. For example, in the C4SL maps most of Greater Glasgow is classified as exceeding the C4SL for only allotments, with a small area (1.5% of total area = 4.7 km²) northwest of the city centre where the garden C4SL is exceeded (Figure 5.8A). However, other maps, such as the transformed unit map (Figure 5.7A), shows more variability in soil Pb classes. Conversely, the neighbourhood maps show heterogeneity in the C4SL maps (Figures 5.8A and B). In these maps there are large areas exceeding the garden C4SL (22% of total area in Paisley (0.04 km²) and 29% of total area in Bishopbriggs (0.05 km²)). Furthermore, there are soil Pb hotspots
where the residential and even public open space C4SL is exceeded (11% of total area in Paisley (0.02 km$^2$) and 6% of total area in Bishopbriggs (0.01 km$^2$)).

Overall, suitable classification breaks are important when communicating the spatial patterns of soil Pb to decision makers. This has been noted for other types of environmental data (e.g. noise pollution (Bartels and Van Beurden, 1998) and flood risk mapping (Fariza et al., 2017)).

5.5.3.2. Conveying uncertainty

The kriging error maps (Figure 5.6) indicate where predicted soil Pb concentrations are less certain. Uncertain predictions are generally located at larger distances from known samples, along study area boundaries, or along natural barriers i.e. the River Clyde. Interestingly, the uncertainty of the Glasgow city predictions and, to a lesser extent, Bishopbriggs predictions are relatively constant (Figures 5.6A and C), reflecting the almost pure nugget models. For Paisley there is decreased uncertainty in the vicinity of measurement points reflecting the small nugget and semivariance over small lags.

Uncertainty is a common property of kriging maps. The uncertainty in predictions can be conveyed using side-by-side maps (Figure 5.6) or transparency (Figures 5.7 and 5.8), with each method having pros and cons. Other methods of conveying uncertainty are also available e.g. annotation, colour hue or saturation, and crispness (MacEachren et al., 2005), but these were not applicable in this instance.

The side-by-side maps (e.g. Figure 5.6) are easily interpreted as they clearly depict where predictions are more uncertain. The side-by-side display was favoured over toggling by decision makers in an urban planning model, according to a web-survey (Aerts et al., 2003). However, the side-by-side display can take up valuable space on the page and divert the reader’s attention from the main message.

Transparency is one of MacEachren et al.’s (2005) new visual variables for depicting uncertainty and is based on the metaphor of fog making our view more unclear. It is more subtle than side-by-side maps, as it requires detailed investigation of the map to locate the most uncertain locations. However, soil Pb’s spatial pattern and its associated uncertainty can be conveyed in one map, which may be preferred by
decision makers. Bisantz et al. (2011) found transparency was an important tool when communicating missile uncertainty to decision makers.

**Figure 5.9:** Kriged soil Pb maps for Paisley using the G-BASE samples (A) and using samples from this study (B). Note different soil Pb classification break points and kriging extents between the maps, but the same geographic scale is used.
Figure 5.10: Kriged soil Pb maps for Bishopbriggs using the G-BASE samples (A) and using samples from this study (B). Note different soil Pb classification break points and kriging extents between the maps, but the same geographic scale is used.
5.5.3.3. Communicating results

Mapping scale can dictate the resolution of kriged maps. For instance, kriged soil Pb maps for Paisley and Bishopbriggs using the G-BASE dataset show little detail (Figures 5.9A and 5.10A). This is because the prediction grid was 200 x 200 m, the mean sample separation distance was 356 m, but the study areas were only 750 m x 300 m. This is compared with the more detailed spatial patterns shown in soil Pb values in the neighbourhood kriged maps (Figures 5.9B and 5.10B). However, in these maps the prediction grid was 5 x 5 m, average sample separation distance was 30 m, and study areas were the same areas. This emphasises the importance of choosing the most appropriate sample density based on study area size.

Similarly, city-wide kriged predictions indicates relatively low soil Pb concentrations across Greater Glasgow i.e. <400 mg kg\(^{-1}\) (Figure 5.7A), as any outliers in the original dataset have been smoothed by kriging. However, kriged maps of the neighbourhood datasets reveal some areas of concerningly high soil Pb concentrations in Glasgow (>520 mg kg\(^{-1}\); Figures 5.7B and C). Therefore, if you were to look at the city scale map you might assume soil Pb levels across Glasgow were relatively low. However, this is clearly not the case. This highlights the importance of mapping scale when kriging soil Pb.

5.4.3.4. End users

The choice of mapping units, classification breaks, uncertainty communication, and mapping scale all depends on the end user of the map. These are important issues that need to be addressed before soil Pb can be spatially interpolated.

Mapping units depend on the output and kriging methodology. Generally, the most accurate predictions are for transformed values, but these have limited practical interpretability, although they illustrate the relative pattern in soil Pb concentrations. Therefore, they could help decided where further investigation into soil Pb is required, or where future remediation should be targeted.

Classification breaks can influence the amount of spatial heterogeneity visible in the kriged maps. They can also be standardised to allow comparison between different locations. The C4SL maps could be useful for end users who want to
decide future land uses i.e. an allotment should not be placed on locations with soil Pb > 80 mg kg$^{-1}$, unless further investigation and remediation is carried out.

The sampling density will be controlled by the study area of the map i.e. does soil Pb need to be mapped at the city or neighbourhood scale? Nested sampling could also be explored, in order to quantify soil Pb variability at smaller scales. Also two-step sampling could be employed e.g. 10% of sample effort focused on the larger scale to identify the most at risk locations, with the remaining 90% of samples collected from these locations.

5.6. Conclusions

This chapter has found that the spatial dependence in soil Pb concentration is only present at samples separated by short lag distances (<1.6 km in Glasgow, and <60 m in Paisley and Bishopbriggs). Therefore, in all datasets close neighbouring points may not significantly improve our understanding of the dataset, so model-based statistical analysis is required when further analysing the datasets. The spatial dependence is not spatially homogenous throughout the datasets. Local indicators of spatial autocorrelation indicates samples with positive spatial dependence, either as hotspots or coldspots, are spatially clustered in each dataset.

The Paisley dataset shows the strongest spatial dependence, perhaps due to more point Pb sources. Soil Pb concentrations vary on the 50 - 120 m scale at the neighbourhood scale, but up to 10 km in the city scale. This difference is due to different sampling intervals; therefore, any future work should carefully consider the sampling strategy before spatially interpolating soil Pb.

Differences in sampling intervals can also be observed in the kriged soil Pb map at the city and neighbourhood scales. As before, each study area contains hotspots and coldspots of soil Pb concentrations. The causes behind the spatial patterns in soil Pb concentrations will be explored in the next chapter by expanding linear mixed model to include covariates.

The display unit, classification breaks, and uncertainty communication method should be carefully considered before producing kriged soil Pb maps. The choice in each will depend on the end user of the map. In this study kriged soil Pb predictions were displayed using Box-Cox transformed units with a side-by-side
kriging error map; transformed units (mg kg\(^{-1}\)) with uncertainty depicted using transparency; and exceedance of C4SLs for different land uses with more uncertain predictions appearing more transparent.

**References - Chapter 5**


Appendix 5.1: Moran’s I R code

# Read in the data (N.B. code shown below is for the G-BASE dataset but it also applies to Paisley and Bishopbriggs):
GBASE.df <- read.csv("G_BASE_complete.csv", header=T)

# Box-Cox transform:
bc <- boxcox (Pb~1, data= GBASE.df, lambda=seq(-1.5,1.5,0.01))
print (lambda <- bc$x[which.max(bc$y)])
GBASE.df $bcPb <- ((GBASE.df $Pb^lambda)-1)/lambda

# Create a spatial weights matrix (W). Use k = 1 to find minimum distance at which all points have a distance-based neighbour:
coordinates(GBASE.df) = ~Easting + Northing # make spatial object.
Pb1_nb <- knn2nb(knearneigh(coordinates(GBASE.df), k = 1))
dsts <- unlist(nbdists(Pb1_nb, coordinates(GBASE.df)))
summary(dsts) #check minimum distance > 1.

# Identify neighbours as any points within k=1 max distance of another point:
d784 <- dnearneigh (coordinates(GBASE.df), min(dsts), max(dsts))
# Calculate the Euclidean distance between each neighbouring point:
dist_784 <- nbdist(d784, coordinates(GBASE.df))
# Calculate inverse distance to a power of 1:
dist.inv_784 <- lapply (dist_784, function(x) 1/x)
# Transform to a list format, style=W ensures values row-standardised:
G.weights <- nb2listw(d784, glist=dist.inv_784, style= "W")

# Map neighbouring points:
plot (coordinates(GBASE.df), pch=19, xlab="Easting", ylab="Northing")
plot (G.weights, coordinates(GBASE.df), add=TRUE, col="red", length=0.08)

# Run Global Moran's I:
Mt <- moran.test (GBASE.df$bcPb, G.weights, randomisation=FALSE, alternative = "two.sided")

# Display expected and observed I values from Monte Carlo simulation:
set.seed(138)
bperm <- moran.mc(GBASE.df$bcPb, listw = G.weights, nsim = 999)
plot (bperm, xlab="", main="")
expect <- as.numeric (mt$estimate[2])
abline (v=(expect), lty=3)
text ((expect+0.05), 5, paste("Observed(I)"))
text (0.33, 5, paste("Expected(I)"))

# Correlogram generation:
sp.cor <- sp.correlogram(d784, GBASE.df$bcPb, order=6, method="I", style="w", randomisation=FALSE)
plot (sp.cor, main=""")

# Local Moran's I (LISA):
Lisa <- localmoran(GBASE.df$bcPb, G.weights, alternative = "two.sided")
# Assess spatial dependence and significance of dataset:
par(mfrow=c(1,2))
hist(lisa[,1], xlim=c(-2,2), breaks=40, col="grey", xlab="Local Moran's I", main="")
hist(lisa[,5], col="grey", breaks=30, xlab="p-value", main="")
# Samples with positive spatial dependence:
length(which(lisa[,1]>=0.3))
# Samples with significant spatial dependence:
length(which(lisa[,5]<=0.05))
# Samples with positive and significant spatial dependence:
length(which(lisa[,5]<=0.05 & lisa[,1]>=0.3))

# Moran's scatterplot:
Create new column for standardised values:
GBASE.df$sbcPb <- scale(GBASE.df$bcPb)
# Create a lagged variable:
GBASE.df$lag_bcPb <- lag.listw(G.weights, GBASE.df$sbcPb)
# Plot lagged vs original data:
signif.df <- subset(GBASE.df, subset = lisa[,5]<=0.05)
bf <- lm (GBASE.df$lag_bcPb ~ GBASE.df$sbcPb)
cf <- round (coef(bf), 3)
r <- round((summary(bf)$r.squared), digits=3)
eq <- paste0("lag soil Pb = ", cf[1], " + ", (cf[2]), " soil Pb")
par(mfrow=c(1,1))
plot (GBASE.df$sbcPb, GBASE.df$lag_bcPb, xlab="Standardised soil Pb (box-cox)", ylab="Lagged and standardised soil Pb (box-cox)")
points (signif.df$sbcPb, signif.df$lag_bcPb, pch=3)
abline (h = 0, v = 0)
abline (bf, lty = 3, lwd = 3, col = "red")
text (1.2, 1, paste("Quad 1 (HH)", col="red", font=2)
text (-1.5, -0.8, paste("Quad 2 (LL)", col="red", font=2)
text (1.2, -0.8, paste("Quad 3 (HL)", col="red", font=2)
text (-1.5, 1, paste("Quad 4 (LH)", col="red", font=2)
text (1.9, -1.2, paste(eq), pos=2, cex=0.8)
text (1.9, -1.4, paste("R2= ", r), pos=2, cex=0.8)

# Map LISA results:
# Assign points to quadrants based on their standardised and lag value and if LISA is significant (i.e. <0.05)
GBASE.df$quad_sig <- NA
GBASE.df@data [(GBASE.df$sbcPb >= 0 & GBASE.df$lag_bcPb >= 0) & (lisa[,5] <= 0.05), "quad_sig"] <- 1
GBASE.df@data [(GBASE.df$sbcPb <= 0 & GBASE.df$lag_bcPb <= 0) & (lisa[,5] <= 0.05), "quad_sig"] <- 2
GBASE.df@data [(GBASE.df$sbcPb >= 0 & GBASE.df$lag_bcPb <= 0) & (lisa[,5] <= 0.05), "quad_sig"] <- 3
GBASE.df@data [(GBASE.df$sbcPb <= 0 & GBASE.df$lag_bcPb >= 0) & (lisa[,5] <= 0.05), "quad_sig"] <- 4
GBASE.df@data [(lisa[,5] >= 0.05), "quad_sig"] <- 5

# Export as .csv so can load into GIS and map
P.Moran.I.results.df <- cbind (GBASE.df$Samp.Numb, GBASE.df$Pb, GBASE.df$Easting, GBASE.df$Northing, GBASE.df$sbcPb, GBASE.df$quad_sig, GBASE.df$lag_bcPb, lisa)
write.csv (GBASE.Moran.I.results.df, "GBASE_moran_i_results.csv")
Appendix 5.2: Comparing semivariogram and kriging methods

**Aim:** determine the most accurate method (and platform) for fitting a semivariogram model and kriging a dataset.

**Methodology:** fit a semivariogram to the Paisley and Bishopbriggs datasets using three different methods. Then krige soil Pb in each study area.

1. **Classical geostatistics in R:**
   a. Optimum lambda values were determined.
   b. Mean soil Pb values were Box-Cox transformed.
   c. A robust (Cressie-Hawkins) estimator was used to generate the empirical semivariogram, as this was shown to perform better than the standard Matérn estimator due to the presence of outliers.
   d. Different bin/lag distance and total distances (i.e. number of lags) were tested, and the option generating the RMSE closest to 1 was selected.
   e. Different models were fitted to the empirical semivariogram using WLS. Two different model types were tested (Spherical and Exponential) and initial covariance values (i.e. the partial sill and the range) were experimented with.
   f. The models were assessed using LOOCV, with the mean and median SSE values calculated.
   g. The resulting models were used to krige the sample points, to estimate soil Pb values for a 5 x 5m grid over the study area.
   h. The kriging error was calculated from the resulting surface.
   i. Kriging predictions and error were exported and plotted in GIS.

2. **Classical geostatistics in GIS:**
   a. The mean soil Pb values from each sample location were transformed using a Box-Cox transformation, with optimal lambda values input.
   b. The empirical semivariogram was generated, as there is no option to change the semivariogram estimator it is assumed that a standard Matérn estimator was used.
   c. The lag distance and number of lags were determined from R analysis, which had already optimised these variables for each dataset.
   d. The empirical semivariogram was then modelled. Exponential and spherical models were tested to correspond to the R analysis. These were optimised using WLS, which finds the optimal nugget, partial sill and range values, which minimise the weighted sum of squares (though this value is not given). For all models a standard neighbourhood search type was used, with two – five neighbours, one sector, and the range as the
search radius. This was to ensure that weights were only applied based on distance i.e. no smoothing factor was added.
e. During model generation there was the option to include measurement error. However, this was simply a percentage, so was not able to vary between points (e.g. some sample locations might have had larger measurement errors than others, based on differences between duplicate measurements). Therefore, this was left as 0% i.e. no measurement error.
f. Lastly, the kriging results were cross-validated for the sample points using LOOCV. These cross-validation statistics were exported so that the mean and median theta errors could be calculated. Determining total kriging error was more difficult.

3. Model-based geostatistics in Matlab:
a. See Sections 5.3.2. and 5.3.3., and Appendix 5.3.

4. Comparing methods:
a. The outputs from each platform were assessed by LOOCV. The mean and median of $\theta(x)$ (theta) were calculated (Equation 5.10). This is comparable between different software platforms (unlike the AIC which relies on a likelihood value that is not consistent).

Results:

- In both Paisley and Bishopbriggs, GIS and R generate similar results as they used very similar methods and parameters (Table 5.7). Unclear if spherical or exponential model is best.
- For Paisley, REML with no measurement error generates the 'best' cross-validation statistics, as the $\bar{\theta}(x)$ is closest to 1 (0.95) and the $\tilde{\theta}(x)$ is closest to 0.455 (0.41). Interestingly; however, if measurement error is included the mean and $\tilde{\theta}(x)$ errors drop off significantly (0.78 and 0.09 respectively). This is because there is more flexibility with these predictions, which are more realistic but give poorer cross-validation statistics. These number may have been better if an independent validation dataset had been used instead of cross-validation.
- A similar pattern can be observed in Bishopbriggs, as REML with no measurement error generates the best $\bar{\theta}(x)$ (1) and $\tilde{\theta}(x)$ (0.48).
- This suggests that REML without measurement error is the best model for both Paisley and Bishopbriggs.
- NB it is difficult to compare kriging errors from different software as it depends on the method used and the how the prediction grid was generated. For example in GIS the kriging error is only available for the known sample points, not all predicted locations, so is much lower as only looks at 111-138 points instead of several thousand.
Table 5.7: Comparison of semivariogram parameters, lags, validation statistics and kriging error from semivariogram models created using different software platforms and models for Paisley and Bishopbriggs. The 'best' values are highlighted in green for validation statistics. Expo. = exponential model, Sph. = spherical model, ME = measurement error. Differs from results in Table 5.5, where lambda is not fixed.

<table>
<thead>
<tr>
<th>Area</th>
<th>Software (method)</th>
<th>PAISLEY</th>
<th>BISHOPBRIGGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model/options</td>
<td>Matlab (REML)</td>
<td>GIS (WLS)</td>
</tr>
<tr>
<td>Semivariogram parameters</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.0014 0.0004</td>
<td>0.0067 0</td>
<td>0.0013</td>
</tr>
<tr>
<td>Partial sill</td>
<td>0.0575 0.054</td>
<td>0.0612 0.0645</td>
<td>0.061</td>
</tr>
<tr>
<td>Range</td>
<td>36.69 37.152</td>
<td>149.78 76.00</td>
<td>92.27</td>
</tr>
<tr>
<td>Smoothness</td>
<td>2.5 2.5</td>
<td>n/a n/a</td>
<td>n/a n/a</td>
</tr>
<tr>
<td>Lambda</td>
<td>-0.25 -0.25</td>
<td>-0.25 -0.25</td>
<td>-0.25 -0.25</td>
</tr>
<tr>
<td>Measurement error</td>
<td>0 0.0031</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lags</td>
<td>Max lag</td>
<td>n/a n/a</td>
<td>350 350</td>
</tr>
<tr>
<td></td>
<td>No lags</td>
<td>n/a n/a</td>
<td>10 10</td>
</tr>
<tr>
<td></td>
<td>Bin width</td>
<td>n/a n/a</td>
<td>35 35</td>
</tr>
<tr>
<td>Validation statistics</td>
<td>Likelihood</td>
<td></td>
<td>720.996</td>
</tr>
<tr>
<td></td>
<td>Min. weight sum of squares</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean $\hat{\theta}(x)$</td>
<td>0.95 0.78</td>
<td>0.92 0.94</td>
<td>0.9 0.91</td>
</tr>
<tr>
<td>Median $\hat{\theta}(x)$</td>
<td>0.41 0.09</td>
<td>0.6 0.68</td>
<td>0.68 0.72</td>
</tr>
<tr>
<td>Kriging error/variance</td>
<td>Mean</td>
<td>0.1412 0.1373</td>
<td>0.0319 0.0265</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>0.1437 0.1394</td>
<td>0.0320 0.0261</td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>1000.63</td>
<td>973.61</td>
</tr>
</tbody>
</table>
Appendix 5.3: Semivariogram and kriging Matlab code

load BB

% Select variable and covariates:
z=meanPb; %use mean soil Pb measurements as variable
n=length(z); %find number of samples
M=ones(n,1); %no covariates selected

% 1. Data preparation:
X = [easting northing]; %Rescale coordinates to make computation simpler
lx = max(X(:,1)) - min(X(:,1)); %range of eastings
mx = [min(X(:,1)) min(X(:,2))]; %minimum easting and northing
X(:,1) = (X(:,1) - mx(1))/lx; %rescale eastings
X(:,2) = (X(:,2) - mx(2))/lx; %rescale northings
k = find(floor(sample)~=sample); %separate duplicates by small distance
X(k,1) = X(k,1) + 0.000001; %separate duplicates by small distance
[X(a,:),] = unique(sample); %check unique
z = z(a); %check unique
M = M(a,:); %check unique
H = distance(X',X'); %create distance matrix

% 2. REML fitting of variogram:
pfixed = NaN*ones(6,1); %create matrix for 6 parameters
pfixed(6) = 0; %no measurement error as complicates results
options = []; %no extra options
[L,par,C,zt,Lm,WI,beta] = ML_fit_opts(pfixed,H,z,M,options);
er = eff_range(par,max(max(H))); %calculate effective range
out_table = table(par(1), par(2), er, par(4), par(5), par(6), L)

% 3. Plot the variogram:
figure (1)
r = 0.01:0.1:350;
c0 = par(1);
c1 = par(2);
range = par(3)*lx;
nu = par(4);
[c]= Cmat(r,c0,c1,range,nu,[],ones(size(r)));
% Method of moments variogram points
zt = bc(z,par(5));
[math,cbin,~,~,~] = mom(H*lx,[],zt,zt,bins,[],[]); % MOM
%Plot
subplot(1,1,1)
p1 = plot(r, c0+c1-c,'k', 'LineWidth', 1.5);
hold 'on'
p4 = plot(cbin, math, 'gx', 'LineWidth', 1.5);
axis([0 350 0 0.006])
xlabel('Lag distance (m)')
ylabel('Variogram (transformed units)')

% 4. Kriging:
g_space = 5; % grid spacing
T = Make_grid(X,g_space/lx,125/lx,0); %grid point included if <125 m of data point
Torig = T;
MT=ones(length(T),1); % no covariates
[out,map_mean,map_sd,sim_out]=predict_options(par,X,T,M,MT,z,options);
% sim out is 1000 simulations at each T site
krige_t=out(:,1); % kriged prediction (transformed)
se_t=sqrt(out(:,2)); % kriging standard error (transformed)
krige=mean(sim_out,2); % kriged prediction
se=sqrt(var(sim_out,[1,2])); % kriging standard error

% Export outputs
UT_x=(Torig(:,1)*lx)+mx(1);
UT_y=(Torig(:,2)*lx)+mx(2);
output = table(UT_x,UT_y,krige_t,se_t,krige,se);

% 5. 10-fold cross validation - WARNING takes ½ hour to run:
nfold = 10; % randomly allocate each measurement one of 10 folds.
fold = ceil(nfold*rand(length(z), 1));

% rerun model without one fold and compare predicted and actual results & repeat 10 times
z_cv = NaN*ones(length(z), 1); % create empty variables to fit
zt_cv = NaN*ones(length(z), 2);
v_cv = NaN*ones(length(z), 1);
for i = 1:nfold;
f = find(fold==i); % allocate fold samples
nf = find(fold~=i); % allocate non-fold samples
[Li,pari,Ci,zt_i,Lmi,WIi,betai]=ML_fit_opts(pfixed,H(nf,nf),z(nf),M(nf,:),options);
[outi,~,~,sim_outi]=predict_options(pari,X(nf,:),X(f,:),M(nf,:),M(f,:),z(nf),options);
% use kriging to predict non-fold values
zt_cv(f,1) = outi(:,1);
zt_cv(f,2) = bc(z(f),pari(5));
v_cv(f) = outi(:,2);
z_cv(f) = mean(sim_outi,2);
end

me = mean(z-z_cv) % Mean error original units
rmse=mean(sqrt((z-z_cv).^2)) % root mean square error original units
me_t = mean(zt_cv(:,1)-zt_cv(:,2)); % Mean error transformed units
rmse_t = mean(sqrt((zt_cv(:,1)-zt_cv(:,2))^2)); % RMSE trans.
msse = mean(((zt_cv(:,1)-zt_cv(:,2))^2)./(v_cv)) % mean squared standardised prediction error should be 1
medsse = median(((zt_cv(:,1)-zt_cv(:,2))^2)./(v_cv)) % mean standardised squared error should be 0.45
sse=(zt_cv(:,1)-zt_cv(:,2))./(sqrt(v_cv));
sum(sse)
out_table=table(me,rmse,msse,medsse)
figure(3)
hist(sse)
title('Bishopbriggs')
xlabel('SSE')
ylabel('Frequency')
Chapter 6. Sources of soil Pb at different scales and locations.

The previous chapter identified that spatial dependence and variation was present in soil Pb concentrations in all study areas. However, the factors controlling Pb addition to urban soil need to be fully understood in order to successfully use covariates to predict and map soil Pb concentration in unsampled locations (Chapter 8). However, the relationship between soil Pb and different covariates may vary with scale and location. This needs to be fully explored before any machine learning commences, in order to decide which covariates should be included in different models and to determine how transferrable the models will be.

6.1. Literature

Multiple studies have explained how variations in soil Pb levels are controlled by different factors or covariates (Table 6.1).

<table>
<thead>
<tr>
<th>Co-variate</th>
<th>Example studies</th>
<th>Explanation</th>
<th>More details at</th>
</tr>
</thead>
<tbody>
<tr>
<td>Land Use</td>
<td>Birke &amp; Rauch, 2000</td>
<td>Usually, highest soil Pb observed in industrial and residential land uses, lowest in agricultural or wooded land. Related to fewer Pb inputs.</td>
<td>Section 6.1.1., Figure 6.1 Table 6.2</td>
</tr>
<tr>
<td></td>
<td>Lark &amp; Scheib, 2013</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Mao et al., 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fordyce et al., 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road distance</td>
<td>Deska et al., 2011</td>
<td>Historical vehicle exhaust (via Pb-petrol) increased soil Pb levels next to the road. Lead levels decrease away from roads.</td>
<td>Section 6.1.2., Figure 6.2 Table 6.3</td>
</tr>
<tr>
<td></td>
<td>Dao et al., 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Wang et al., 2022</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Traffic density</td>
<td>Fakayode &amp; Olu-Owolabi, 2003</td>
<td>Roads with higher traffic densities pollute surrounding soil with higher Pb concentrations and for a greater distance.</td>
<td>Section 6.1.3., Table 6.3</td>
</tr>
<tr>
<td></td>
<td>MacKinnon et al., 2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Historic industry</td>
<td>Kelly et al. 1996</td>
<td>Some types of historic industry generated Pb emissions and/or Pb waste which were deposited to nearby soils. Depends on industry age, type, distance, and density.</td>
<td>Section 6.1.4., Tables 6.4 and 6.5</td>
</tr>
<tr>
<td></td>
<td>Pacyna &amp; Pacyna, 2000</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Glennon et al., 2014</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>McIlwaine et al., 2017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building age</td>
<td>Thornton et al., 1985</td>
<td>Soils around older buildings have higher Pb levels. Possibly due to more Pb-paints used, but Pb could also originate from additional sources e.g. ash tipping, petrol, and past industry.</td>
<td>Section 6.1.5., Figures 6.3 - 6.5</td>
</tr>
<tr>
<td></td>
<td>Davies &amp; Thornton, 1986</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Davies et al., 1987</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Yesilnisan et al., 2008</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Building distance</td>
<td>Haar &amp; Aronow, 1974</td>
<td>Pb-paint sources in buildings have a greater impact closer to building foundations.</td>
<td>Section 6.1.6., Figure 6.6.</td>
</tr>
<tr>
<td></td>
<td>Schwarz et al., 2012</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Soil type</td>
<td>Lacarce et al., 2012</td>
<td>Soil landscape influences Pb levels, as the pH, soil texture, and OM content influences Pb mobility. Parent material has some influence on soil Pb levels but is less important in urban areas.</td>
<td>Section 6.1.7., Figure 6.7</td>
</tr>
<tr>
<td></td>
<td>Johnson et al., 2017</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.1.1. Land use

Land use alone is not a direct control on soil Pb concentrations but combines multiple other sources of Pb. Therefore, the relationship between soil Pb and land use is generally proven at the broad scale. For example, there tends to be higher soil Pb concentrations closer to urban centres (Birke and Rauch, 2000; Mao et al., 2014; Li et al., 2014). Li et al. (2014) attributed higher soil Pb levels within the inner ring road in Beijing to a longer history and higher density of traffic, compared with urban soil beyond the outer ring roads. Similarly, Birke and Rauch (2000) detected elevated Pb levels in Berlin’s inner-city soils. These levels decreased away from the city centre and were likely due to historic traffic and coal emissions from domestic and industrial sources. However, other Pb sources e.g. Pb paint and past industrial activity, can disrupt this pattern. Mao et al. (2014) demonstrated that in Beijing land use only has a significant effect on Pb concentrations within the transitional zone (i.e. between the city centre and suburbs), whereas in the suburbs and urban core land use type did not influence soil Pb levels.

The lowest Pb concentrations tend to occur in woodland and agricultural soils, compared with other land uses (e.g. Gibson and Farmer, 1983; Birke and Rauch, 2000; Fordyce et al., 2012, 2017; Lark and Scheib, 2013; Table 6.2). These lower Pb levels are likely due to woodlands and agricultural land being predominantly located on the city periphery, thus are further from many inner-city Pb sources.

![Conditional boxplots of soil Pb concentrations for different land uses in Dublin, note soil Pb values are log transformed (From Glennon et al., 2014).](image)
<table>
<thead>
<tr>
<th>Land Use</th>
<th>N.</th>
<th>Mean</th>
<th>S.D.</th>
<th>Median</th>
<th>Min.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Glasgow</strong> - Gibson and Farmer (1983)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>397</td>
<td>216</td>
<td>211</td>
<td></td>
<td>22</td>
<td>1900</td>
</tr>
<tr>
<td>Domestic Gardens</td>
<td>80</td>
<td>244</td>
<td>191</td>
<td>37</td>
<td>1310</td>
<td></td>
</tr>
<tr>
<td>Parks</td>
<td>103</td>
<td>226</td>
<td>190</td>
<td>22</td>
<td>1260</td>
<td></td>
</tr>
<tr>
<td>Open Spaces</td>
<td>169</td>
<td>219</td>
<td>244</td>
<td>23</td>
<td>1900</td>
<td></td>
</tr>
<tr>
<td>Agricultural</td>
<td>45</td>
<td>136</td>
<td>122</td>
<td>24</td>
<td>749</td>
<td></td>
</tr>
<tr>
<td><strong>Berlin</strong> - Birke and Rauch (2000)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inside city</td>
<td>2183</td>
<td>119</td>
<td>77</td>
<td></td>
<td>4710</td>
<td></td>
</tr>
<tr>
<td>Outside city</td>
<td>1564</td>
<td>39</td>
<td>28</td>
<td></td>
<td>469</td>
<td></td>
</tr>
<tr>
<td>High density res.</td>
<td>476</td>
<td>133</td>
<td>109</td>
<td></td>
<td>1490</td>
<td></td>
</tr>
<tr>
<td>Industrial</td>
<td>459</td>
<td>117</td>
<td>87</td>
<td></td>
<td>4710</td>
<td></td>
</tr>
<tr>
<td>Low density res.</td>
<td>826</td>
<td>73</td>
<td>50</td>
<td></td>
<td>2070</td>
<td></td>
</tr>
<tr>
<td>Woodland</td>
<td>614</td>
<td>34</td>
<td>26</td>
<td></td>
<td>269</td>
<td></td>
</tr>
<tr>
<td>Agricultural</td>
<td>388</td>
<td>30</td>
<td>25</td>
<td></td>
<td>307</td>
<td></td>
</tr>
<tr>
<td><strong>London</strong> - Lark and Scheib (2013)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>6245</td>
<td>297</td>
<td>381</td>
<td>184</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Industrial</td>
<td>64</td>
<td>226</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Domestic garden</td>
<td>1554</td>
<td>196</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road verge</td>
<td>597</td>
<td>172</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urban open space</td>
<td>1119</td>
<td>162</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>195</td>
<td>161</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agricultural¹</td>
<td>760</td>
<td>108</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woodland</td>
<td>274</td>
<td>86</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Beijing</strong> - Mao et al. (2014)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All</td>
<td>457</td>
<td>22</td>
<td></td>
<td>&lt;1</td>
<td>292</td>
<td></td>
</tr>
<tr>
<td>Industrial</td>
<td>15</td>
<td>42</td>
<td>8</td>
<td>292</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Roadside</td>
<td>101</td>
<td>25</td>
<td>2</td>
<td>93</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential</td>
<td>76</td>
<td>24</td>
<td>4</td>
<td>82</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institutional</td>
<td>38</td>
<td>20</td>
<td>7</td>
<td>69</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Park</td>
<td>55</td>
<td>17</td>
<td>5</td>
<td>32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forest</td>
<td>55</td>
<td>20</td>
<td>4</td>
<td>52</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Conversely, the highest Pb concentrations are often observed in industrial soils (Birke and Rauch, 2000; Lark and Scheib, 2013; Glennon et al., 2014; Fordyce et al., 2017; Table 6.2). This is due to atmospheric deposition of industrial Pb emissions, combined with pollution from industrial waste and coal burning emissions (see Section 6.1.4.). However, the type and age of industry is important. For instance, Glennon et al. (2014) found the highest soil Pb levels in areas of Dublin associated with heavy industry e.g. dockyards and metal works, compared with light industry and other land uses (Figure 6.1).

Residential soil can also have elevated soil Pb concentrations, compared with other land uses (Fordyce et al., 2012; Lark and Scheib, 2013). This could be due to unique Pb sources in gardens, such as Pb-paint, burning of household waste, and spreading of coal ash (see Section 6.1.5.). Interestingly, Birke and Rauch (2000) observed higher soil Pb levels in high-density residential soil, compared with soil around low-density housing (Table 6.2). They explain that high-density buildings are more likely to be located in the inner-city, where other Pb sources are/were more prevalent.

Overall, there is a large variability in soil Pb levels within a single land use (Table 6.2). Gibson and Farmer (1983) found non-agricultural land uses had very similar mean soil Pb concentrations, and large standard deviations indicating high variability within land uses (Table 6.2). This is because most land uses do not have direct Pb sources, but a combination of factors.

Johnson and Ander (2008) warn that results from different studies should not be directly compared if the studies used different sampling schemes, sampling depths (dilution effect), and analytical methods. The studies may also vary in the statistical methodology used i.e. some studies transform the data, and the selection and definition of land use classes vary. Overall, results between studies are very difficult to compare, but land use classes can be ranked from highest Pb to lowest Pb to give a relative estimate of similarity.
6.1.2. Road distance

The addition of Pb to petrol in the 20th century emitted millions of tons of Pb into the atmosphere, thereby elevating Pb concentrations in roadside soils worldwide (Steinberg, 2002; Oudijk, 2010). Numerous studies have examined the relationship between soil Pb and traffic emissions and have observed exponentially decreasing soil Pb with increasing distance from a road (Figure 6.2; Table 6.3).

This pattern of decreasing soil Pb concentrations away from roads can be easily overprinted by other Pb sources in urban soils, which is why elevated soil Pb levels in roadside soils are not always observed. For instance, Farmer and Lyon (1977) measured soil Pb concentrations in parks in Glasgow and observed higher soil Pb levels around the park peripheries i.e. <5 m from a road, compared with park interiors (Table 6.3). In a later Glasgow wide study, Gibson and Farmer (1983) found higher soil Pb concentrations in open spaces <10 m from a road (166 mg kg$^{-1}$), compared with spaces >10 m from a road (99 mg kg$^{-1}$). Similarly, in Baltimore Yesilonis et al. (2001) observed 3.5-fold higher Pb levels in soils within a 100 m buffer around roads, compared with soils outside the road buffer. Lastly, in Appleton and Cave’s (2017) study of London, soil Pb concentrations only decreased away from roads in areas which had never been urbanised.

Table 6.3 illustrates that the range and extent of Pb pollution around roads varies between studies and is likely influenced by numerous factors. Firstly, soil properties can influence the extent of Pb elevation in roadside soils. Kluge and Wessolek (2012) detected elevated Pb topsoil concentrations 2.5 m from the road, but soil solution Pb concentrations were not raised until 10 m from the road edge. Elevated Pb levels in the embankment soil (<2.5 m from road) correlated with an increased sorption capacity due to a high pH from calcareous materials, and high organic carbon levels from tyre wear.
<table>
<thead>
<tr>
<th>Study</th>
<th>Traffic Density (vpd)</th>
<th>Land Use (Location)</th>
<th>Road Age</th>
<th>Sampling Interval (m)</th>
<th>Road Distance (m)</th>
<th>Pb (mg kg⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gibson and Farmer, 1983</td>
<td>Various</td>
<td>Urban (Glasgow)</td>
<td>-</td>
<td></td>
<td>&lt;10 m</td>
<td>166*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>&gt;50 m</td>
<td>99*</td>
</tr>
<tr>
<td>Fakayode and Olu-Owolabi, 2003</td>
<td>Low (&lt;6000 vpd)</td>
<td>Urban (Osogbo, Nigeria)</td>
<td>5, 10, 30,</td>
<td>5</td>
<td>22*</td>
<td></td>
</tr>
<tr>
<td></td>
<td>High (&gt;36,000 vpd)</td>
<td></td>
<td>50</td>
<td></td>
<td>10</td>
<td>16*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...50</td>
<td>11*</td>
</tr>
<tr>
<td>Deska et al., 2011</td>
<td>Medium (13,000 vpd)</td>
<td>Agricultural (E. Poland)</td>
<td>-</td>
<td>0, 50, 100</td>
<td>0</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td>24</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>100</td>
<td>21</td>
</tr>
<tr>
<td>Bai et al., 2011</td>
<td>Low</td>
<td>Agricultural (Guizhou Plateau, SW China)</td>
<td>1, 5, 10, 20, 50, 100, 200, 250</td>
<td>1</td>
<td>76.75</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>36.75</td>
</tr>
<tr>
<td>MacKinnon et al., 2011</td>
<td>Low</td>
<td>Agricultural (W. Glasgow)</td>
<td>0.6, 1.5, 3, 5, 10, 20, 30, 50</td>
<td>0.6</td>
<td>703</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>453</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>251</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...50</td>
<td>95</td>
</tr>
<tr>
<td>Yan et al., 2012</td>
<td>Low</td>
<td>Agricultural (Tibetan Plateau)</td>
<td>-</td>
<td>0, 10, 50</td>
<td>0</td>
<td>28*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>23*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>50</td>
<td>21*</td>
</tr>
<tr>
<td>Dao et al., 2014</td>
<td>High (25,000 vpd)</td>
<td>Urban park (Dublin)</td>
<td>19th C, 5 (grid)</td>
<td>0</td>
<td>498*</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>128*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td>94*</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>...70</td>
<td>37*</td>
</tr>
<tr>
<td>Kluge et al., 2014</td>
<td>High (100,000 vpd)</td>
<td>Urban embankment (Berlin)</td>
<td>1921, 2.5</td>
<td>2.5</td>
<td>245</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>208</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7.5</td>
<td>147</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>800</td>
<td>24</td>
</tr>
<tr>
<td>Wang et al., 2022</td>
<td>High (73,000 vpd)</td>
<td>Urban park (Sydney)</td>
<td>1921, 5</td>
<td>20</td>
<td>725</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>25</td>
<td>702</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>30</td>
<td>655</td>
</tr>
</tbody>
</table>
The sampling strategy may also create disparities between studies measuring soil Pb levels away from a road (Table 6.3). Airflow and runoff dynamics cause petrol emissions to elevate soil Pb concentrations only for the first few centimetres to meters from the road edge, with background concentrations usually reached by 20 - 50 m (Harrison and Johnston, 1985; Kluge et al., 2014). Therefore, the coarse sampling intervals in some studies likely obscured small scale changes in Pb’s distribution. For example, Deska et al. (2011) collected samples at 50 m intervals, so witnessed a sharp decrease between soil Pb levels immediately next to the road and 50 m from the road edge. Even studies which collected sample at a 5 m interval noted a rapid drop off in soil Pb concentrations (Figure 6.2; Dao et al., 2014). The studies which best captured the fine scale spatial variability in soil Pb concentrations from a road where those which staggered their sampling interval by collecting more samples closer to the road (MacKinnon et al., 2011).

Vegetation can also influence roadside soil Pb levels because trees or shrubs lining roads intercept Pb emissions, thereby limiting Pb deposition to the soil (Zhang et al., 2012). Pb distribution is also dependent on the prevailing wind direction, which elevates Pb concretions on downwind side of roads, if there are no obstacles (Zhang et al., 2012; Walraven et al., 2014). Overall, Pb levels in roadside soils are elevated up to ~20 m away from a road, although this distance depends on numerous factors. The highest Pb concentrations occur in soils immediately next to (<5 m) the road edge, which is predominantly due to the splashing and runoff of relatively mobile Pb from vehicle exhaust or roadside dust, combined with atmospheric deposition (Harrison and Johnston, 1985; Kluge et al., 2014). At greater distances from the road i.e. >20 m, atmospheric deposition of Pb dominates (Harrison and Laxen, 1981).

Owing to the ban of leaded petrol in the UK in 1999, Pb emissions from vehicles are predominantly historical. However, contemporary vehicles still emit small amounts of Pb. This primarily arises as brake pads and tyres contain 57 and 11 mg kg\(^{-1}\) of Pb respectively (Apeagyei et al., 2011). Subsequent wear of these materials will emit some Pb to the atmosphere, an estimated 32 tonnes in the UK in 2010 (Kummer et al., 2009). Additionally, Pb is present in petrol; an average 17µg of Pb is emitted per litre of unleaded petrol (Kummer et al., 2009). Other sources of Pb in roadside soils includes the erosion of Pb-chromate (CrO\(_4\)Pb) yellow road markings, which contain an average of 13% wt. Pb (Adachi and Tainosho, 2004).
6.1.3. Traffic density

Another important difference between the studies in Table 6.3 is the traffic density of the roads around which the soil Pb levels were measured. Generally, higher density roads pollute roadside soils to a greater degree and for a further distance. For example, Fakayode and Olu-Owolabi (2003) observed higher Pb concentrations 5 m from roads with a high traffic density compared with low density roads in Nigeria. Furthermore, they found that soil Pb concentrations level off at around 10 m distance from low density roads but are elevated to at least 30 m from high density roads. In the UK, Chamberlain et al. (1979) found that 22% of Pb emissions were deposited <100 m from a motorway. However, in urban areas this distance is significantly decreased due to the masking effect from the urban background Pb concentration of 0.5-1 µg/m³ created by the total road network. Harrison and Johnston (1985) attributed higher atmospheric Pb deposition rates next to a motorway in summer, with higher traffic densities during these months.

Traffic density is not the only constraint on soil Pb levels, as driving conditions have proven to be an important determinant in Pb emissions. Specifically, more heavy metals are emitted in stop-start conditions i.e. when vehicles rapidly accelerate, decelerate, or idle, compared with when they maintain a steady speed (Ewen et al., 2009). Therefore, soil Pb emissions levels tend to be highest around congested inner-city roads, especially if the road has two or more traffic signals and/or a roundabout (Charlesworth et al., 2003; Aslam et al., 2013). Lead emissions in urban streets can also be intensified by the canyon effect (Ewen et al., 2009).

Higher Pb emissions along inner city roads were also found by Gibson and Farmer (1981), who observed higher atmospheric tetraalkyl and particulate Pb in Glasgow city centre (96 ng/m³) compared with a rural site (3.9 ng/m³). They also noted that levels were higher during the day (53 ng/m³) than at night (11 ng/m³) due to higher traffic densities and slower moving traffic. Additionally, they observed a lower-than-expected decrease in atmospheric Pb with increasing height above the street (36% reduction at 12 m), which is caused by poor ventilation and the urban canyon effect.

The difference between urban and rural traffic emissions is reflected in the studies in Table 6.3, in that agricultural roadside soils generally have lower Pb levels than urban soils. This is because Pb in urban soils will likely originate from other sources (Preciado et al., 2007), with isotopes often used to differentiate between petrol and other Pb sources (see Chapter 7).
6.1.4. Historic industry

Historical industrial processes have elevated soil Pb concentrations via deposition of atmospheric emissions and/or industrial waste. Different types of historical industry emitted varying levels of Pb (Table 6.4). Thus, Pb levels in surrounding soils will depend on the type(s) of industry nearby (summarised in Table 6.5).

In cities multiple different industries were often clustered together. Therefore, quantifying Pb pollution from the closest historical industry may be futile, as there could be overprinting from multiple point sources and general background emissions. Consequently, historical industry density may have a stronger influence on soil Pb concentrations. For example, Glennon et al. (2014) used the locations of historic industries in Dublin to generate a spatial kernel density estimation (SKDE) which had a strong positive correlation with soil Pb levels ($R^2=0.45$).

Relatively recent Pb emission data are outlined in Table 6.4. However, emission values will have been higher in the past, i.e. before the decline in British heavy manufacturing industries and the introduction of pollution control measures in the past 70 years (e.g. the Control of Pollution Act of 1974 and the Environmental Protection Act of 1974). Consequently, the age of the industry may influence the amount of Pb pollution generated, but exact values are difficult to quantify.

There has been limited generalised research into Pb emissions from historical industrial processes, due to clear difficulties in obtaining data and large variations within industrial processes (Table 6.5). However, several studies have noted decreasing soil Pb concentrations away from a specific type of industry or point source (e.g. a Pb mine (Moles et al., 2004); a battery factory (Skinner and Salin, 1995); a cement plant (Lin et al., 2007); and a glass factory (Rybick and Kyzio, 1987); Table 6.5). Wind and slope direction can often impact the distribution and range of soil Pb pollution. However, these studies tend to focus on rural areas where there is less overprinting of Pb from multiple industries and/or other sources.
Table 6.5: Industrial processes producing Pb, an explanation of why they generate Pb, the type of Pb pollution produced, and some example case studies where possible (based on UNEP (2010a), Alloway (1995) and Burt (1984)). Grouped by industrial type.

<table>
<thead>
<tr>
<th>Group</th>
<th>Industry type</th>
<th>Explanation</th>
<th>Material</th>
<th>Pb (mg kg(^{-1}))</th>
<th>Location</th>
<th>Reference</th>
<th>Pollution mechanism</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metals</td>
<td>Ore mining</td>
<td>Mining of galena, cerussite, and anglesite extracts Pb, but waste material can contain high Pb levels.</td>
<td>Soil</td>
<td>Max. = 80,000</td>
<td>Newtownards, NI</td>
<td>Moles et al., 2004</td>
<td>- Waste tailings</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Soil</td>
<td>Mean = 6,902</td>
<td>Wanlockhead, S. Scotland</td>
<td>Moffat, 1989</td>
<td>- Gravitational transport</td>
</tr>
<tr>
<td></td>
<td>Smelting Pb</td>
<td>Smelting removes impurities in ore but produces slag and coarse textured (&gt;5 μm) dust containing Pb. Recently factories started recycling Pb from dust and slag.</td>
<td>Soil</td>
<td>Mean = 437</td>
<td>Pb smelter, Port Pirie, Aus.</td>
<td>Tiller et al., 1976</td>
<td>- Wind depo.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Soil</td>
<td>Max. = 35,300</td>
<td>Pb smelter, Pribram, CZ</td>
<td>Ettler et al., 2004</td>
<td>- Slag</td>
</tr>
<tr>
<td></td>
<td>Iron (Fe) and steel production</td>
<td>Pb and Fe are closely related in some ore deposits, so industrial processes to produce Fe releases Pb. Steel further heats Fe to burn out impurities.</td>
<td>Dust on tree bark</td>
<td>Max. = 30,580</td>
<td>Scrap steel plant, France</td>
<td>Geagea et al., 2008</td>
<td>- Dust</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Marine sediment</td>
<td>Mean = 85</td>
<td>Next to ship-breaking area, Bangladesh</td>
<td>Siddique et al., 2012</td>
<td>- Fugitive emissions</td>
</tr>
<tr>
<td></td>
<td>Shipyards</td>
<td>Paint with &gt;30% Pb was used to coat ships to prevent corrosion. Lead was released to atmosphere from thermal cutting, painting, paint removal, welding, chipping, and burning of steel during ship overhaul and breaking.</td>
<td>Soil</td>
<td>4,232 - 5,733</td>
<td>Chittagong ship breaking</td>
<td>Hossain &amp; Islam, 2006</td>
<td>- Runoff to marine sediments</td>
</tr>
<tr>
<td>Fossil fuels</td>
<td>Coal Mining</td>
<td>Lead adsorbs strongly to OM which was later compressed to form hydrocarbons. Scottish coals average Pb = 24 mg kg(^{-1}). Acidic conditions cause oxidation and leaching of Pb.</td>
<td>Coal gauge</td>
<td>21-25</td>
<td>Coal mining area, Huainan, China</td>
<td>Fang et al., 2015</td>
<td>- Coal gauge</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Soil</td>
<td>Max. = 27</td>
<td></td>
<td></td>
<td>- Coal sludge</td>
</tr>
<tr>
<td></td>
<td>Coal combustion</td>
<td>Heating coal volatilises Pb it contains, which is emitted in fly ash. These ultra-fine particles can be transported long distances (&lt;1000 km). In the furnace bottom ash particles collect, which are coarser and concentrate Pb levels.</td>
<td>Fly ash</td>
<td>Mean = 24</td>
<td>Indian coal</td>
<td>Tiwari et al., 2014</td>
<td>- Coal dust</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Bottom ash</td>
<td>Mean = 117</td>
<td></td>
<td></td>
<td>- Fly ash, atmospheric deposition</td>
</tr>
<tr>
<td></td>
<td>Petrol production</td>
<td>Tetraethyl-lead and tetramethyl-lead manufacturing mixed molten Pb and sodium (Na). During the process high levels of Pb were emitted, and waste contains Pb.</td>
<td>Garden soil</td>
<td>Max. = 1,200</td>
<td>Near Tetraethyl-lead manufacturing plants</td>
<td>Mench et al., 2012</td>
<td>- Bottom ash, waste or infill</td>
</tr>
<tr>
<td>Industry</td>
<td>Description</td>
<td>Site</td>
<td>Author(s)</td>
<td>Notes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------</td>
<td>-------------------</td>
<td>----------------------------------------</td>
<td>---------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Railways</td>
<td>Railway lines traditionally used coal as a fuel source, and mineral railways will have transported, and spilled, industrial material possibly high in Pb. Atmospheric dispersion of particulate Pb means that soil Pb levels tend to decrease exponentially away from railway sidings.</td>
<td>Soil, Mean = 273</td>
<td>In cleaning bay, railway junction, Poland</td>
<td>Maławski &amp; Wilkomirski, 2001</td>
<td>Dust via coal combustion, spillages, industrial waste as fill material</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cement production</td>
<td>Industrial waste e.g. fly/bottom ash or Pb-zinc (Zn) smelter sludge, were sometimes used in cement production, because they are a source of silica and Fe, have pozzolanic properties, are a fuel source.</td>
<td>Soil, Mean = 497</td>
<td>In railway storage yard</td>
<td>Ge et al., 2000</td>
<td>Dust</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brick-making</td>
<td>Fireclay (mudstones under coal seams) and high clay industrial waste products e.g. coal gangue, and fly/bottom ash, were often used to produce low quality bricks. Volatile metals in the industrial waste or fireclay may be emitted during firing.</td>
<td>Soil, Mean = 72</td>
<td>500 m from cement plant</td>
<td>Lin et al., 2007</td>
<td>Dust</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brick-making</td>
<td>Fireclay (mudstones under coal seams) and high clay industrial waste products e.g. coal gangue, and fly/bottom ash, were often used to produce low quality bricks. Volatile metals in the industrial waste or fireclay may be emitted during firing.</td>
<td>Soil, Mean = 31</td>
<td>Around a brick plant, Hualian, China</td>
<td>Zhou et al., 2014</td>
<td>Dust</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dye works</td>
<td>Lead compounds e.g. Pb-acetate and litharge (PbO) are/were used as dyes and mordants in dyeing and calico-printing yellows, greens, and oranges. High levels of water are required so Pb can occur in high levels in wastewater.</td>
<td>Wastewater, Max. = 19 mg L⁻¹</td>
<td>Current dyeworks, India</td>
<td>Sivakumar et al., 2011</td>
<td>Wastewater (leaching)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glass production</td>
<td>Adding Pb to glass lowers the melting point, improves viscosity, and increases density and the refractive index. Some forms of Pb, e.g. Pb antimonite (Pb₃Sb₂O₇) with 58 wt.% Pb, were traditionally used to colour glass. The Pb levels vary with glass type, e.g. optical glass contains up to 50 wt.% PbO, and leaded crystal must contain &gt;24 wt.% PbO, with up to 90 wt.% PbO. During glass manufacturing the furnace reach 1650 °C, so some PbO will be volatised to the atmosphere.</td>
<td>Soil, 538</td>
<td>50 m from glassware plant, Inowroclaw, Poland</td>
<td>Rybicki &amp; Kyzio, 1987</td>
<td>Dust, Waste</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ceramic production</td>
<td>Lead compounds with up to 50 wt.% Pb e.g. PbO₃, and red/white-Pb, have been used in pottery glazes for centuries. They lower the glaze’s melting point, increase the firing range, fill in defects, and produce a glossy finish. Ceramic production can volatise up to 10% of Pb and produce Pb in waste material.</td>
<td>Soil, Mean = 6,420 Max. = 20,000</td>
<td>Near a pottery kiln, Mexico</td>
<td>Hibbert et al., 1999</td>
<td>Dust, Waste</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chemical</td>
<td>Pigments and paint production</td>
<td>Lead-pigments were added to paint as white-Pb provides a thick 'hiding' coat, red-Pb reduces corrosion, and Pb naphthenate/oxalate are drying agents. Wastewater and sludge, containing high Pb levels may be leached to nearby soil.</td>
<td>Soil</td>
<td>Max. = 33,100</td>
<td>Former paint factory, New Jersey, USA</td>
<td>UESPA, 2015</td>
<td>- Sludge</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Battery production</td>
<td>Lead-acid batteries are used to power motor and electrical vehicles. They are made from Pb alloys which are coated with black powder containing 60-80 wt.% PbO. ~8 kg of Pb is emitted to the atmosphere per 1,000 batteries produced.</td>
<td>Soil</td>
<td>Mean = 2,000</td>
<td>Battery factory, Ibadan, Nigeria</td>
<td>Onianwa &amp; Fakoyode, 2000</td>
<td>- Dust</td>
<td>- Waste</td>
</tr>
<tr>
<td>Match and firework production</td>
<td>Low levels of Pb dioxide are combined with red phosphorus as an oxidising powder in matchstick production. Pb dioxide and nitrate were used as oxidising agents and to give the 'cracking' effect in fireworks production.</td>
<td>Soil</td>
<td>Mean = 20</td>
<td>Match factory, Peshwar, Pakistan</td>
<td>Zaeem et al., 2014</td>
<td>- Dust</td>
<td></td>
</tr>
<tr>
<td>Pesticide production and use</td>
<td>Lead arsenate [Pb₃OH (AsO₄)] pesticides were applied to apple orchards until the 1960s to reduce moth and insect populations.</td>
<td>Park soil</td>
<td>Max. = 10,297</td>
<td>Near a fertiliser plant, Galway, Ireland</td>
<td>Carr et al., 2008</td>
<td>- Waste</td>
<td></td>
</tr>
<tr>
<td>Munitions production</td>
<td>Lead levels in shooting or firing range soils are elevated due to the degradation of Pb bullet fragments. Production of ammunition also elevates Pb levels in the surrounding soil, mainly due to the dumping of ammunition waste, but also from wastewater discharge.</td>
<td>Soil</td>
<td>Mean = 37,000</td>
<td>Burning site, explosive factory, Scotland</td>
<td>Ellis et al., 2002</td>
<td>- Waste</td>
<td>- Wastewater</td>
</tr>
<tr>
<td>Waste incinerators, destructors, and tips</td>
<td>Municipal solid waste includes paints, glass, and batteries, which are high in Pb. On average 300 mg kg⁻¹ of combustible Pb is present in the waste. The incinerator heats the municipal solid waste to &gt;850°C which will volatise some Pb to atmospheric particles and fly ash (15%), with the remaining Pb (85%) concentrated in the residual bottom ash.</td>
<td>Soil</td>
<td>Mean = 350</td>
<td>Municipal solid waste incinerator, Newcastle</td>
<td>Rimmer et al., 2006</td>
<td>- Dust</td>
<td>- Fly ash</td>
</tr>
<tr>
<td>Printing</td>
<td>Pre-computerised printing methods used type casts which were commonly an alloy of Pb (50-90 wt.%), tin (Sn) and antimony (Sb). These casts were moulded, cut, and eroded, which generated Pb dust which may have been tracked outdoors by shoes or clothing. Later linotypes were used, which melted Pb-alloys at 300°C to produce plates of text.</td>
<td>Street dust</td>
<td>Range = 5,000 to 200,000</td>
<td>&lt;150 m from newspaper printers, London</td>
<td>Jensen &amp; Laxen, 1985</td>
<td>- Dust</td>
<td></td>
</tr>
</tbody>
</table>
6.1.5. Building age

The Environmental Geochemistry research group at Imperial College London first noted the close relationship between building age and soil Pb levels in British urban areas in the 1980s. This was based on Thornton et al.’s (1985) 1981-82 national survey of 53 British towns and cities, where they measured Pb levels in household dust, road dust, playground dust, and garden soil. The main aim of their study was to determine the range of Pb exposure in urban environments. However, it also discovered higher Pb concentrations in garden soils of houses built before 1880, compared with houses constructed post-1950 (Figure 6.3). They attributed lower Pb levels around post-1950s houses with a reduction of Pb in paint, and a shorter history of fossil fuel residues.

These findings prompted Davies and Thornton (1986) to undertake a more detailed investigation into housing age and soil Pb in Brighton. They sampled garden soil along with household and road dust and ensured that five housing age categories were evenly sampled. The results (Figure 6.4) show increasing soil Pb levels with increasing housing age, and sharp drop in Pb levels from different aged houses in Brighton. Values are mean Pb levels based on 20-38 samples. Based on Davies and Thornton (1986).
after 1919, which could be linked with the ban on interior Pb-paints in 1922. Furthermore, road dust shows a similar trend, suggesting that the two pools interact. However, the study noted that older houses are more commonly located in the city centre, where there is more traffic and congestion.

A similar study was conducted in Birmingham and found pre-1952 houses had significantly higher garden soil Pb than post-1952 houses, 547 mg kg\(^{-1}\) and 393 mg kg\(^{-1}\) respectively (Davies et al., 1987). More recently, Appleton and Cave (2018) found that areas of London which were urbanised before 1940 contained approximately twice as much soil Pb as areas which were developed after 1940. They also found that this pattern was consistent for several other UK cities (e.g. Coventry, Derby, Sheffield, and Swansea). Overall, these four studies indicate that in the UK soil Pb is elevated around houses built before 1920 to 1950.

One possible explanation for the link between soil Pb and housing age could be paint. Pb pigments, such as white-Pb \([2\text{PbCO}_3\cdot\text{Pb(OH)}_2]\), red-Pb \([\text{Pb}_3\text{O}_4]\), and Pb-chromate \([\text{PbCrO}_4]\), have been added to paints for centuries (Nriagu, 1983). Specifically, white-Pb increases the thickness and durability of the paint, red-Pb reduces corrosion, and Pb napthenate and oxalate are drying agents (Thornton et al., 2001). Therefore, paints often contained 40-70% wt. Pb (Inskip and Hutton, 1987).

The health dangers associated with Pb-based paints have been recognized to painters since the 18th century (Penrose, 2000). However, Lockhart Gibson first linked Pb-paint and Pb poisoning in 1904. Since then numerous studies have linked Pb-paints with elevated blood Pb levels (e.g. Clark et al., 1985; Mielke and Reagan, 1998; Tagne-fotso et al., 2016; Pavilonis et al., 2022). Lead contained within Pb-paints enters the bloodstream via direct ingestion of Pb-paint chips, inhalation of Pb dust created during renovation of Pb-painted surfaces, and ingestion of soils containing Pb from paint sources (Marcus and Elias, 1995). The last two pathways interact, as there is a strong correlation between Pb levels in indoor dusts and the amount of Pb in the soil (Davies et al., 1987). This is due to demolition, renovation, or burning of Pb painted surfaces, which adds Pb to the soil via settling of Pb-rich dusts (Inskip and Hutton, 1987). Farfel et al. (2003) found that Pb levels in dust increased 40 times during the demolition of older (pre-1950) buildings due to high Pb levels in paint. Peeling and flaking of poorly maintained external Pb-painted surfaces also contributes Pb to soil levels (Turner and Sogo, 2012).
The use of Pb-based paint varies globally, as legislation reducing and banning the addition of Pb to paint differs between countries. For example, The League of Nations banned the addition of white Pb to interior paints in 1922 (Barnett et al. 2006). However, the USA only reduced Pb levels in interior paint the 1950s and banned it in 1976 (Barnett et al. 2006). These dates were earlier in some states i.e. Baltimore banned interior Pb-paints in 1951 (Markowitz and Rosner, 2013). Lead in exterior paints was not reduced until the 1970s in the USA, with the Pb-paint poisoning prevention act in 1971 (Lin-Fu, 1992).

Consequently, many North American studies have observed higher Pb levels around buildings constructed before 1960-78. The USEPA public guidance suggests that homes built before 1940 have an 87% chance of containing Pb-paint, which decreases to 69% for 1940-1959 houses, and 24% for 1960-1977 homes (USEPA, 2021). These dates loosely match other studies, for instance Jacobs et al. (2002) found that houses built before 1960 had a five to eight times greater hazard from Pb-paints, compared with post-1960 homes. Similar results were found for Baltimore, with higher soil Pb levels occurring next to buildings constructed between 1920 and 1930 (Figure 6.5: Schwarz et al., 2012). Furthermore, a strong negative correlation between the year of construction (1860-2000) and topsoil Pb concentrations ($R^2=0.7$) was also observed in Baltimore (Yesilonis et al., 2008).

![Figure 6.5: Soil Pb concentrations (ppm) around houses of different ages in Baltimore. Dotted black line shows when Pb-paint was used (1884-1978) and solid line shows when Pb-petrol was in used (1929-1986). The grey line is the USEPA limit for soil Pb (400ppm). From Schwarz et al., 2012.](image-url)
The UK was slightly quicker in reducing Pb levels in paint, banning white interior Pb-paints in 1926 and labelling all paints with >1% wt. Pb with a warning from 1963 (Johnson et al., 1984). An EU regulation in 1992 (Council Directive 89/677/EEC) caused the UK to ban the addition of Pb-pigments to all paints, except from those for specialist use (UNEP, 2010b). A DEFRA (2005) pamphlet suggests that Pb-paint is likely present in British houses constructed before the mid-1960s, although warns that Pb-paint was around until the 1980s.

Lead-paint may not be the only source of Pb in the soil surrounding older buildings. Before gas central heating became commonplace in the 1970s, coal fires were predominantly used for heat generation, and coal stoves were used for cooking (Appleton et al. 2012). The coal remains were often disposed of in gardens to increase soil fertility, especially in inner city areas (Thornton, 1991; Meuser and Van de Graaff, 2011; Fordyce et al., 2012). However, Scottish coals contains on average 24 mg kg\(^{-1}\) of Pb and can hold up to 137 mg kg\(^{-1}\) of Pb (Farmer et al., 1999). The low temperatures reached in domestic fires meant that most of the Pb would have accumulated in the coal ash (Meuser, 2010). Subsequently, constantly adding this coal ash to garden soils will have gradually increased Pb levels.

Sewage sludge was also historically added to domestic gardens and allotments to increase fertility (Chaney et al., 1984; Meuser, 2010). Sewage sludge is composed of by-products of municipal and industrial waste, which can contain high levels of Pb (mean = 820 mg kg\(^{-1}\); Berrow and Webber, 1972; Alloway, 1995). An estimated 0.59-0.99 kg of Pb per ha could have been added to domestic gardens each year, for over 30 years (Harrison et al. 1975). This constant addition of Pb will have raised soil Pb levels. This was indicated in Leicester, where soil Pb concentrations of up to 385 mg kg\(^{-1}\) were observed in gardens where sewage sludge had been spread for 74 years (Bridges, 1991). Regulations restricting the amount of Pb allowed in sewage sludge to <2000 mg kg\(^{-1}\) were introduced in the early 1980s (Harrison and Laxen, 1981).
6.1.6. Building distance

Soil Pb levels have also been proven to be higher closer to older buildings. For example, Haar and Aronow (1974) observed higher topsoil Pb concentrations within 0.6 m of a building, compared with 3 m from buildings in Detroit. Clark and Knudsen's (2013) high scale sampling of 71 properties in Wisconsin found that soil Pb concentrations decreased away from the drip line of the house to the nearby road, and this relationship was stronger in pre-1960s buildings. Similarly, Schwarz et al. (2012) proved that soil Pb levels decrease away from buildings in Baltimore, and this decrease largely occurs within the first 5 m from the building (Figure 6.6). Likewise, Jordan and Hogan (1975) found soil Pb levels decreased with increasing distance from wooden houses in Christchurch, New Zealand.

The relationship between building distance and soil Pb levels is difficult to prove, as many other factors can obscure it. For instance, Pb-petrol emissions can add to soil Pb concentrations, if gardens are small and next to a road. Demolition and rebuilding can also influence the relationship (Clark and Knudsen, 2013), and sampling can be limited by the amount of soil surrounding a building, with higher Pb concentrations in bare soil compared with grassy lawns (Thornton et al., 1985).

Another factor controlling the amount of Pb in soils surrounding buildings is the building material. Multiple studies have proven that painted, wooden homes release more Pb to the environment than brick houses. Harr and Aronow (1974) observed that soils 0.6 m from houses with painted frames had mean Pb levels of 2,349 mg kg\(^{-1}\), which was over six times higher than soil 0.6 m from brick houses in Detroit (mean = 351 mg kg\(^{-1}\)). Jordan and Hogan (1975) found similar results for Christchurch, with soil next to wooden houses ranging from 400-7,500 mg kg\(^{-1}\) compared with soil near brick houses measuring 20-85 mg kg\(^{-1}\). This is likely due to the higher amounts of Pb within painted houses, and additionally the greater release of Pb from these surfaces. Davis et al. (2001) discovered that on average five to 50 times more Pb is released per m\(^2\) of wall surface area for painted wood (520 µg) compared with unpainted wood (9 µg) and brick (94 µg).
6.1.7. Soil type

Soil type is linked to Pb levels in two ways, firstly through the weathering of Pb from the underlying parent material, and secondly from higher Pb retention in certain types of soils.

Many studies have linked parent material and soil Pb levels (e.g. Yang et al., 2009; Lacarce et al., 2012), although these tend to be conducted in rural areas or regions with mineralised soils. For instance, a study by BGS and DEFRA found that in England the normal background concentration of Pb in soils is significantly higher in regions with mineralised soils (background concentration = 2,400 mg kg$^{-1}$), compared with urban (background concentration = 820 mg kg$^{-1}$) and rural (background concentration = 180 mg kg$^{-1}$) soils (Figure 6.7; Ander et al., 2013). Furthermore, parent material has a greater influence on Pb levels in lower soil horizons, whereas anthropogenic Pb sources tend to dominate in the upper horizons (Walraven et al., 2013a; Bing et al., 2014). Although this pattern may be clear in rural soils, it is often disturbed or non-existent in urban soils due to substantial soil alterations (Mcllwaine et al., 2017).

Figure 6.7: Maps showing (A) Pb domains and (B) interpolated topsoil Pb concentrations in England (From Ander et al., 2013).

Parent material more often has an impact on soil properties which can influence Pb mobility. Typically, Pb migrates slowly to the mineral horizons, so can reside in temperate soils for up to 20,000 years (Kylander et al., 2008; Stille et al., 2011).
Initially, Pb exhibits low mobility due to its affinity with OM in the organic (O) horizon, where Pb is strongly bound to ligands or other negative functional groups e.g. carboxylic and humic acids (Gangloff et al., 2014; Chotpantarat et al., 2015). Decomposition of OM causes leaching of Pb to the mineral horizons where it is complexed with Fe and Mn(hydr)oxides e.g. ferrihydrite and birnessite, via inner sphere absorption (Kaste et al., 2006; Schroth et al., 2008). Additionally, clay rich soils have a high cation exchange capacity, so have slower Pb migration rates than sandy soils (Walraven et al., 2013b). The above relationships are pH dependant, as studies on coniferous forest soils have proven that Pb mobility is higher in acidic conditions, due to Pb desorption from complexes and/or the dissolution of Pb minerals e.g. pyromorphite (Stille et al. 2011).

Overall, Pb has higher mobility in sandy, acidic soils with low organics. Therefore, these types of topsoils will have experienced greater Pb leaching, and consequently exhibit lower Pb levels. These factors have caused topsoil properties to be input as covariates in several predictive models. For example, Lacarce et al. (2012) used soil texture as a covariate to predict soil Pb and found that fine textured soils had higher Pb levels. Similarly, Johnson et al. (2017) input soil landscape as a controlling factor in their linear model and observed that areas of waterlogged soil had the biggest impact on soil Pb levels, as swampy areas will have a more acidic pH and higher OM levels.

### 6.2. Chapter aims

The aims of this chapter are:

1. To determine which covariates influence soil Pb concentrations in Greater Glasgow, Paisley, and Bishopbriggs.

2. To assess if the same factors constantly influence soil Pb concentrations at different locations and scales.

3. To explore how including covariates can improve modelling soil Pb and representation in map form.
6.3. **Methodology**

The literature analysis identified that seven key factors, henceforth called covariates, have previously correlated with soil Pb concentrations (Table 6.1). These were: land use, road distance, traffic density, historic industry, building age, building distance, and soil type. However, historic industry can be split into historic industry type, distance, age, and density. Similarly, soil type includes OM content, pH, and soil texture. Therefore, a total of 12 covariates were explored.

The relationship between soil Pb and each covariate was first examined in the Greater Glasgow dataset, as this contains more data points. The results from this initial study were then used to narrow down the covariates, to avoid the issue of multiple hypothesis testing in the smaller neighbourhood datasets (Lark, 2017). In Paisley and Bishopbriggs only seven potential covariates were investigated. These were: land use, road distance, building distance, building age, historic industry distance, historic industry type, and soil organic matter percentage. Some of these covariates are categorical while others are continuous, meaning the methodology varies slightly.

6.3.1. **Covariate determination**

Each covariate in question (i.e. the 12 covariates at the city scale and the seven covariates at the neighbourhood scale) were predominantly determined using a nearest neighbour join in QGIS (Version 3.16.11).

At the city scale, land use was classified by the ordnance survey (OS) MasterMap Greenspace layer™ (OS, 2022), which uses OS MasterMap topographic data and aerial photography to classify urban greenspaces into a land use class (Figure 6.8A). This layer was spatially joined with the location of G-BASE samples using QGIS. The OS MasterMap layer has the benefit of covering a whole city. Land use classes were aggregated, as including too many individual land use classes risks over-fitting the model (see Appendix 6.1). At the neighbourhood scale, land use was classified based on field observations and digitisation of aerial imagery.

Road distance and type (Figures 6.8B) were found at all scales using a nearest neighbour join in QGIS. This calculated the distance between each sample point and the nearest road (from OS MasterMap®; Crown copyright, 2004), plus a 3 m buffer to account for total road area. Additionally, it provided the sample point with the attribute of the nearest road type which was aggregated into high density
(motorways), medium density (A, B, and primary roads), and low density (local streets and minor roads). The best measure of traffic density is the number of vehicles per day; however, this information is not freely accessible for all roads. Therefore, road type (metadata on OS MasterMap®) was used as a proxy.

Historic industry covariates were assessed at both the city and neighbourhood scales by digitisation of the footprints of any historic industries falling <40 m from the boundary of the study area, using historic maps ranging from 1864 to 2000. Along with its location, the industry age (Figure 6.8C) was recorded as the first historic map the industry appeared on, and the industry type (Figure 6.8D) was also noted. The historic industry polygons were joined with the sample points using a nearest neighbour join in QGIS. If a sample was >40 m from a historic industry polygon it was classed as non-industrial. This value was selected from literature as the average range of industrial Pb pollution (Table 6.5). At the city scale the 44 past industry types were aggregated into three main groups: light historic industry (LHI), heavy historic industry (HHI), and non-industrial (NHI; see Table 6.6 in Section 6.4.1.4.). There was no aggregation of historic industries at the local scale due to a more limited variability in historic industry types.

Historic industry density was only calculated at the city scale, due to limited data at the local scale. Glennon et al.’s (2014) method of a SKDE of historic industry density in Dublin was adapted slightly to generate a KDE for all historic industries in Greater Glasgow. Firstly, the centroid of each historical industry polygon was calculated, along with its coordinates, using QGIS. The historic industry centroids and Glasgow G-BASE sample locations were loaded in CrimeStat® (Version 3.3; Levine, 2010). Where a single KDE, with a normal interpretation and a fixed interval bandwidth of 500 m, was used to determine the absolute density in units per km² of historic industry around each G-BASE sample point. Other parameters were explored, especially for the bandwidth, but 500 m was selected as this generates a sufficient amount of detail to observe local patterns while not over smoothing the KDE (Figure 6.8E).

At the city scale, three distinctive building age periods were selected, based on the availability of historic OS maps, and previous literature (i.e. pre-1920, 1920-1970, post-1970, Figure 6.8F; see Table 6.8, Section 6.4.1.8. for details). At the neighbourhood scale more specific building age classes were used, depending on the history of the area and the available OS maps. For each time period the OS
historical map closest to the end of the period was used to digitise any buildings (including gardens) present, which were absent in the earlier map. Therefore, the earliest building age was recorded, not the current building age. This is important, as the presence or demolition of older buildings may add Pb to nearby soil. The sample points were then joined with the building age polygons using a nearest neighbour join in QGIS. Any point <5 m (to account for error) from a building polygon was given that polygon’s building age. If a sample point was >5 m from a building polygon it was given the age 0 i.e. no building age.

Soil properties, i.e. OM content, pH, and texture or type, were determined from the collected samples. Soil OM content was calculated using LOI (Figure 6.8G, see Section 2.2.3.3 for details). pH was found via a pH probe, with the methodology detailed in Section 2.2.3.2 (Figure 6.8H). Soil texture was recorded during sample collection using the feel method to determine if the soil was gravel, sand, sandy-clay, sandy-silt, silt, silty-clay, silty-sand, clay, clayey-sand, or clayey-silt (Fordyce et al., 2012). These categories were aggregated into the predominant soil texture i.e. sand (and gravel), silt, or clay (Figure 6.8I).
Figure 6.8A - F: Maps of covariate distribution in the Greater Glasgow study area.
Figure 6.8G - L: Maps of covariate values at sample points in the Greater Glasgow study area.
Figure 6.9: Maps of covariate distribution in the Paisley study
Figure 6.10: Maps of covariate distribution in the Bishopbriggs study area. NB OM and Pb classes match classes in Paisley.
6.3.2. Geostatistical analysis

Standard regression and correlation analyses require the assumption that the residuals from the regression model are identically and independently distributed. This assumption would be appropriate if the measured Pb values were spatially correlated or if the measurement locations were selected at random. The surveys considered in this thesis use a non-random design (Chapter 3) and in Chapter 5 there is evidence of spatial dependence amongst the Pb measurements. Therefore, geostatistical models must be used to account for this spatial correlation when assessing the relationships between Pb measurements and the various covariates. As discussed in Chapter 5, model-based geostatistics is more accurate at spatially interpolating soil Pb, versus classical geostatistics (Appendix 5.2).

The same LMM as used in Chapter 5 (Equation 5.6) was expanded to include covariates. In particular, in the design matrix ($\mathbf{M}$) a column consisted of constant values, for continuous covariates such as soil OM. Alternatively columns of binary values were included in $\mathbf{M}$ to represent the presence or absence of categorical covariates, such as a specific land use class (Marchant, 2018). With the number of columns equalling the number of available classes plus one.

Validation statistics were run to assess model accuracy. Firstly, 10-fold cross validation was performed (see Section 5.3.3. for details). Additionally the Akaike Information Criterion (AIC; Akaike, 1973) was calculated. The best model will have the lowest AIC value, as it is a compromise between the maximised likelihood value ($L$) and the number of model parameters ($k$) (Marchant, 2018):

$$
\text{AIC} = 2k - 2L .
$$

(6.1)

The relationship between soil Pb (Box-Cox transformed) and each covariate was also assessed visually using box-plots of modelled soil Pb values for categorical variables, and scatterplots for continuous variables. In the box-plots the red centre line represents the median ($Q_2$), and the coloured boxes show the inter-quartile range (IQR) i.e. the 25th and 75th percentiles ($Q_1$ and $Q_3$). These are notched to represent the variability of the medians ($Q_2 \pm 1.57(Q_3 - Q_1)/\sqrt{(n)}$), with notches which do not overlap having significantly different medians (95% confidence level). whiskers are the non-outlier minimum and maximum values, with outliers (i.e. any sample $>1.5$ IQR from the box edge) shown as red crosses.
Wald tests (Lark and Cullis, 2004) were conducted to determine if there is a significant difference in soil Pb values between the classes or values of the covariate. In terms of a linear mixed model, these Wald tests are used to assess whether the regression coefficients are sufficiently large or different to signify significant relationships in a manner that accounts for the spatial correlation amongst the data. The null hypothesis is of the form:

$$\mathbf{L}\hat{\beta} = 0,$$

where $\mathbf{L}$ is a matrix of contrasts and $\hat{\beta} = (\beta_1, ..., \beta_q)^T$ is the vector of estimated fixed effects regression coefficients. When a continuous covariate is included in the linear mixed model (i.e. when the fixed effects consist of a constant, $\beta_1$, and a term proportional to the covariate with coefficient $\beta_2$) then a contrast matrix $\mathbf{L} = [0 \ 1]$ will lead to a null hypothesis that $\beta_2$ is zero. When a categorical covariate with $q$ categories is included in the model (i.e. the $\beta_i$ are the mean response for each category) then a contrast matrix (for $q = 4$):

$$\mathbf{L} = \begin{bmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{bmatrix},$$

will lead to a null hypothesis that the $\beta_i$ are all equal and $\mathbf{L} = [-1 \ 0 \ 1 \ 0]$, will lead to a null hypothesis that $\beta_1 = \beta_3$. The null hypothesis states that the Wald statistic ($W$):

$$W = (\mathbf{L}\hat{\beta})^T (\mathbf{L}\hat{\beta}) \mathbf{C} \mathbf{L}^T,$$

where $\mathbf{L}\hat{\beta}$ is a contrast matrix and $\mathbf{L}\hat{\beta} \mathbf{L}^T$ an inverse covariance matrix of the contrast (Lark and Cullis, 2004), will be realized from a chi-squared ($\chi^2$) with degrees of freedom equal to the rank (i.e. the number of linearly independent rows) of $\mathbf{L}$. This generates a $p$-value, so that the null hypothesis (i.e. there is no variation in soil Pb between different classes or values of the covariate) can be rejected with 95% confidence if $p<0.05$, and 99% confidence if $p<0.01$.

To visually explore the influence of the covariates on the spatial pattern of soil Pb, the soil Pb values were predicted at unknown sample points using nearby soil Pb values and values of a spatially correlated covariate(s). Unknown sample points were a 5 by 5 m or 200 by 200 m prediction grid covering the study area, with covariate data obtained via nearest neighbour joins in QGIS (Version 3.16.11). Soil Pb values at each point on this grid were estimated using the E-BLUP (Stein, 1999), as when combined with REML it is a robust model-based geostatistical tool for
spatial prediction in the presence of trend due to a covariate (Lark et al., 2006). An alternative to E-BLUP is universal kriging, regression kriging, or kriging with external drift; however, these models use the classical geostatistical method-of-moments to fit the variogram, which does not account for random spatial variance. The E-BLUP uses the variance parameters from the REML fit variogram to estimate the fixed ($\hat{\tau}$) and random ($\hat{\mu}_p$) effects, which are input into the equation for determining the predicted value at an unknown point (Lark et al., 2006):

$$\hat{z}(f_p) = M_p^T \hat{\tau} + \hat{\mu}_p.$$  (6.5)

Where soil Pb concentrations at unknown location (in this case grid points) are predicted by combing its relationship with a covariate i.e. the fixed effects ($M_p^T \hat{\tau}$), and spatially dependent random effects ($\hat{\mu}_p$) (Lark et al., 2006).

All analysis was undertaken in Matlab (R2020a, MathWorks Inc.®) with a summary of the methodology shown in Figure 6.11 and example code in Appendix 6.2.

![Flow diagram summarising the methodology for comparing soil Pb and covariates, with inputs, process, outputs, and results shown.](image)

**Figure 6.11:** Flow diagram summarising the methodology for comparing soil Pb and covariates, with inputs, process, outputs, and results shown.
6.4. Results and discussion

6.4.1 City scale results - Greater Glasgow

The relationship between soil Pb and each covariate in the Greater Glasgow dataset is given in Table 6.6, along with covariate details and model validation statistics.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Land use</th>
<th>Road</th>
<th>Historic Industry</th>
<th>Building</th>
<th>Soil</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Dist. (log)</td>
<td>Type/ T. Density</td>
<td>Dist. (log)</td>
<td>Type</td>
</tr>
<tr>
<td>Mean</td>
<td>-Road V</td>
<td>14.25</td>
<td>Low</td>
<td>96.17</td>
<td>HHI</td>
</tr>
<tr>
<td>Median</td>
<td>-Garden</td>
<td>14.25</td>
<td>Low</td>
<td>96.17</td>
<td>HHI</td>
</tr>
<tr>
<td>Minimum</td>
<td>-Comm.</td>
<td>0.90</td>
<td>Med.</td>
<td>0.09</td>
<td>LHI</td>
</tr>
<tr>
<td>Maximum</td>
<td>-Res OS</td>
<td>1.50</td>
<td>High</td>
<td>18.63</td>
<td>NHI</td>
</tr>
<tr>
<td>Skew</td>
<td>-Open S</td>
<td>4.26</td>
<td>Natural</td>
<td>N/A</td>
<td>N/A</td>
</tr>
</tbody>
</table>

A - Summary statistics¹/group options

B - Model validation

C - Relationship with soil Pb

\[ Wald \text{ value} = 10.80, p = 0.055, \text{ Wald rank } = 6 \]

\[ Wald \text{ value} = 6.84, p = 0.032*, \text{ Wald rank } = 7 \]

\[ Wald \text{ value} = 3.84, p = 0.057, \text{ Wald rank } = 8 \]

\[ Wald \text{ value} = 11.91, p < 0.01**, \text{ Wald rank } = 14 \]

\[ Wald \text{ value} = 21.37, p < 0.01**, \text{ Wald rank } = 14 \]

\[ Wald \text{ value} = 41.59, p < 0.01**, \text{ Wald rank } = 14 \]

\[ Wald \text{ value} = 2.28, p < 0.01**, \text{ Wald rank } = 14 \]

\[ Wald \text{ value} = 506.78, p < 0.01**, \text{ Wald rank } = 14 \]

\[ Wald \text{ value} = 1.21, p < 0.01**, \text{ Wald rank } = 14 \]

\[ Wald \text{ value} = 13.39, p < 0.01**, \text{ Wald rank } = 14 \]
6.4.1.1. Land use

There is a weak relationship between soil Pb and land use, as the Wald value is >10 but the p value is not quite significant for a 95% confidence interval (p=0.055; Table 6.6C).

Domestic gardens have the highest mean soil Pb levels (Figure 6.12), which has also been witnessed in other research (e.g. Birke and Rauch, 2000; Lark and Scheib, 2013). Elevated soil Pb levels in garden soils is likely due to multiple Pb sources, including weathering of old Pb-based paint, historic leaded petrol emissions, and dumping of coal ash remains (likely containing some Pb) to act as a fertiliser.

Interestingly, natural land has the 2nd highest median soil Pb concentration (Figure 6.12). Natural land uses often occur on the city periphery, thus are further from Pb sources so tend to have lower soil Pb concentrations (e.g. Lark and Scheib, 2013; Fordyce et al., 2017). However, Pouyat et al. (2002) found highest soil Pb concentrations in forests within New York city, due to their high OM levels. Samples with a natural land use do have slightly higher average OM levels than the rest of the dataset (natural samples mean OM = 11%, all samples mean OM = 9%), but this is the smallest group so has high variability (as seen by the wide IQR; Figure 6.12).

![Box-plots of soil Pb (Box-Cox) for each land use class in the Greater Glasgow G-BASE dataset. Summary statistics for each group are given in the table below, with ranks for median soil Pb values.](image-url)

<table>
<thead>
<tr>
<th>Land use</th>
<th>Road verge</th>
<th>Domestic garden</th>
<th>Commercial</th>
<th>Res. open space</th>
<th>Open space</th>
<th>Natural</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>330</td>
<td>194</td>
<td>168</td>
<td>453</td>
<td>122</td>
<td>18</td>
</tr>
<tr>
<td>Median</td>
<td>2.97</td>
<td>3.02</td>
<td>3.01</td>
<td>2.97</td>
<td>2.99</td>
<td>3.01</td>
</tr>
<tr>
<td>Rank</td>
<td>6</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Mean</td>
<td>2.98</td>
<td>3.02</td>
<td>3.00</td>
<td>2.98</td>
<td>2.99</td>
<td>3.00</td>
</tr>
</tbody>
</table>

*Figure 6.12: Box-plots of soil Pb (Box-Cox) for each land use class in the Greater Glasgow G-BASE dataset. Summary statistics for each group are given in the table below, with ranks for median soil Pb values.*
Commercial samples have the 3rd highest mean soil Pb levels (Figure 6.12). Commercial is a broad land use class but includes current industrial sites which may have generated excess soil Pb. (Residential) open space has medium and varying soil Pb levels (Figure 6.12), which is likely due to the influence of other Pb sources causing localised Pb contamination hotspots. For example, in Gibson and Farmer’s (1983) study of soil Pb in Glasgow, they observed that in parks soil Pb was elevated close (<10 m) to roads (mean = 166 mg kg⁻¹), compared with soil >50 m from roads (mean = 99 mg kg⁻¹). Although these samples were collected when leaded petrol was still in use.

Lastly, road verges have the lowest mean soil Pb level (Figure 6.12). This may be due to the sampling strategy (i.e. aggregating five samples over 20 m² area) masking the fine scale variation in roadside soil Pb levels, or that the traffic volume was not high enough on all roads sampled to generate Pb pollution.

6.4.1.2. Road distance

There was no relationship between road distance and soil Pb in the Glasgow dataset, as proven by the very low Wald value (Wald = 0.01), and a corresponding p value close to 1 (p = 0.9; Table 6.6C). Furthermore, there is a wide scatter when road distance (log) and soil Pb (Box-Cox) are plotted (Figure 6.13).

The distances to road for the sample points was highly skewed (skew = 4.3) and therefore this covariate was log transformed. A small constant (0.01) was added to each distance prior to the log transform so that the logarithm could be calculated for sample points with a 0 m road distance because they occurred within the 3 m buffer. This leads to a number of samples having a log road distance of -4.2, and these values have a wide range of soil Pb values (Figure 6.13).

It is likely that the sampling methodology masked any fine scale variation in soil Pb around roads. Specifically, one G-BASE sample is an aggregate of five sub-samples collected over a 20 x 20 m area. Therefore, a nearby road would have varying distances to different sub-samples.
Overall, it is likely that other Pb sources have overprinted the relationship between soil Pb and road distance e.g. historic industrial emissions. Fordyce et al. (2017) and Lark and Scheib (2013) also suggested this when studying the Clyde Basin and London datasets. This is why studies investigating the relationship between roads and soil Pb have collected samples along transect in a small homogeneous area e.g. a park (Dao et al., 2014), thereby reducing the influence of Pb from other sources. Lead isotopes can be used to help identify only Pb concentrations originating from petrol (Chapter 7).

6.4.1.3. Road type

Another explanation for the unexpectedly weak relationship between soil Pb and road distance in the Glasgow G-BASE dataset, could be that some roads have low traffic densities so only produced minimal Pb emissions. The previous analysis was conducted on each point’s proximity to all roads.

Traffic density has a weak relationship with soil Pb levels in the Glasgow G-BASE dataset, as the Wald p value is just <0.05 (p=0.03) and the Wald value is 7, (rank=7; Table 6.6C). However, this may be a case where correlation does not mean causation, as traffic density and soil Pb have an unexpectedly negative relationship. Specifically, samples close to high traffic density roads have the lowest mean soil Pb concentration (although mean values are similar; Figure 6.14). This negative relationship may be due to the sampling strategy, overprinting from other Pb sources, and/or soil properties.

However, there are also uneven group sizes, with almost 50% of samples closest to a road with low traffic density, resulting in the largest range in this group (Figure 6.14). High density roads are under-represented (n = 30), so more samples are required from nearby these road types in order to fully understand the effect of traffic density on soil Pb.

<table>
<thead>
<tr>
<th>Traffic density</th>
<th>Medium</th>
<th>Low</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Count</td>
<td>179</td>
<td>1076</td>
<td>30</td>
</tr>
<tr>
<td>Median</td>
<td>2.99</td>
<td>2.97</td>
<td>2.92</td>
</tr>
<tr>
<td>Rank</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Mean</td>
<td>2.98</td>
<td>2.97</td>
<td>2.89</td>
</tr>
</tbody>
</table>

Figure 6.14: Box-plots of soil Pb (Box-Cox) for each traffic density class in the Greater Glasgow G-BASE dataset. Summary statistics for each group are given in the table, with ranks for median soil Pb values.
6.4.1.4. Historic industry type

The overall low Wald value (3.8) and just non-significant p value (0.057) suggests there is no relationship between soil Pb and historical industry type. This is also reflected in similar mean Pb concentrations for each industrial group (Table 6.7). This may arise from problems with the grouping process, as some types of historic industry may have generated high soil Pb pollution but were grouped with non-Pb polluting industries. However, there are very few examples of each type of historic industry in the dataset, so some aggregation is required. Overall, it is likely that only a very few types of historic industry will have generated Pb pollution, and this may not be consistent from site to site. Furthermore, some industries cause diffuse pollution over a relatively wide spatial area whereas this analysis focuses on localised pollution.

<table>
<thead>
<tr>
<th>Main Group</th>
<th>Sub-group</th>
<th>Industry types</th>
<th>Explanation</th>
<th>N</th>
<th>Soil Pb.bc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Light historic Industry (LHI)</td>
<td>Fossil fuels and by products</td>
<td>Brick/tile/cement, factory (not related to metal or chemical), coal pit/collery/shaft, electrical, gas, mills, oil, power station, sawmill/wood, textile mill, warehouses, general works.</td>
<td>All used some sort of fossil fuel which will contain Pb and may be emitted as atmo. emissions.</td>
<td>234</td>
<td>Mean: 3.00 Med: 2.99</td>
</tr>
<tr>
<td></td>
<td>Transport</td>
<td>Airport, railway line/ sidings/ stations /storage yards, tram/ bus/ railway depot.</td>
<td>Transportation fuel may have generated Pb emissions.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Heavy historic industry (HHI)</td>
<td>Chemical</td>
<td>Bleach/dye works, chemical works, glass works, fireworks production, paint works, pottery, rubber works, scouring works, soap/candle works.</td>
<td>Pb compounds used in chemical industries may be emitted as dust, sludge, or wastewater.</td>
<td>128</td>
<td>Mean: 3.03 Med: 3.02</td>
</tr>
<tr>
<td></td>
<td>Metal</td>
<td>Engineering/shipbuilding, copper, falcon, forge, foundry, galvanising, iron/steel foundry, metal, munitions, ore mining (shafts), scrap, smithy.</td>
<td>Metal working/ mining involving Pb will generate dust and slag with high Pb levels.</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Waste</td>
<td>Clay/gravel/sand pits, old quarries, refuse destructor, sewage works, tip.</td>
<td>Old pits / quarries were often used to dispose of waste, which may contain Pb products, which can leach into the surrounding soil.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Non-industrial (NHI)</td>
<td>Sample &gt;40 m from historic industry polygon.</td>
<td></td>
<td></td>
<td>923</td>
<td>2.97 (2.98)</td>
</tr>
</tbody>
</table>
6.4.1.5. Historic industry distance

The distance to the nearest historic industry polygon also does not have a significant relationship with soil Pb, as the Wald value is 0 (lowest of all covariates) and clearly non-significant ($p = 0.96$; Table 6.6C). This may occur because this is the distance to all types of historic industry, and the previous section indicates that only specific type of historic industry may have generated Pb pollution. Slope and wind direction may also influence the relationship between historic industry distance and soil Pb. Furthermore, as with previous distance analyses i.e. road distance, the fine scale variation is likely masked by the sampling strategy and there are numerous 0 m (-4.2 log) distances that have a large range of soil Pb values (2.05 - 3.86; Figure 6.15).

6.4.1.6. Historic industry age

The earliest date of the historic industry does have a significant influence on soil Pb levels, as the Wald value is medium (11.91) and significant ($p=0.018$; Table 6.6C). This based on samples $<$40 m from a historic industry polygon, and when industry dates have been aggregated into pre-1920, 1920-1970, 1970-2000, and unknown/non-industrial groups. These groups match the building age categories, and are based on individual Wald tests of more numerous industry date categories. The box-plots and individual Wald tests indicate that historic industry constructed before 1920 has significantly higher mean soil Pb values than industry built after 1920 (individual Wald=8.57, $p=0.0034^{**}$; Figure 6.16). This is to be expected as older industries were subject to less stringent environmental laws, making Pb pollution more likely. Furthermore, older industries may have generated Pb pollution over a longer time span.
6.4.1.7. Historic industry density

Soil Pb (Box-Cox) and historic industry density (log) also have a strong positive correlation in the Glasgow G-BASE dataset, as indicated by a high (22.52) and significant (p = <<0.01) Wald value (Table 6.6C). The scatterplot shows a clear positive trend, with greater soil Pb levels as industry density increases (Figure 6.17). This relationship might be stronger if only heavy historic industries were used generate a KDE. However, historic industries are often located in the city centre, which has high soil Pb levels due to a longer history of more intense leaded petrol emissions and domestic coal burning, as noted by Glennon et al. (2014) and Fordyce et al. (2017).
6.4.1.8. Building age

The high and significant Wald value (41.6, p=<0.01) indicates that soil Pb and building age have a strong relationship. Furthermore, the boxplots (Figure 6.18) and individual Wald tests (Table 6.9) show that soil around older (pre-1920) buildings has higher Pb levels than soil around newer (post-1970) buildings (Test A). This difference is weaker i.e. the Wald value is lower, if pre-1920 buildings are compared with 1920-70s buildings (Test B).

<table>
<thead>
<tr>
<th>Table 6.8: Selected building age periods, with justifications of why each period was selected, along with the maps used to digitise their extent.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Period</strong></td>
</tr>
<tr>
<td>---------------------</td>
</tr>
<tr>
<td>Pre-1920</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>1920-1970</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Post-1970</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Higher soil Pb levels around older buildings may arise from a number of reasons. Firstly, old Pb-based paints used to coat the exterior of older buildings could weather into the soil. Secondly, before central heating became common in the 1970s coal ash remains from domestic fires where often deposited onto garden soil to act as fertiliser and will have contained some unvolatised Pb. Lastly, soil around older buildings will have experience a longer history of Pb emissions from leaded petrol and/or historic industry. Older houses also tend to be located in the inner city (Figure 6.8F) which has a higher density of roads and historic industries, especially pre-clean air acts, thus Pb emissions will have been higher here compared with the suburbs.

Soil away from a building has similar soil Pb levels to soil around 1920-70 buildings (Test E), otherwise it is significantly different from pre-1920 and post-1970 buildings (Tests D and F; Table 6.9). The similarity between mean soil Pb away from buildings (2.94) and around 1920-70s buildings (2.98) is possibly due to general atmospheric deposition away from point sources of Pb pollution.
Table 6.9: Individual Wald tests results for building age group.

<table>
<thead>
<tr>
<th>Test</th>
<th>Group A</th>
<th>Group B</th>
<th>Wald</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Pre-1920</td>
<td>Post-1970</td>
<td>39.38</td>
<td><strong>&lt;0.01</strong></td>
</tr>
<tr>
<td>B</td>
<td>Pre-1920</td>
<td>1920-70</td>
<td>9.87</td>
<td>0.0017**</td>
</tr>
<tr>
<td>C</td>
<td>1920-70</td>
<td>Post-1970</td>
<td>18.96</td>
<td><strong>&lt;0.01</strong></td>
</tr>
<tr>
<td>D</td>
<td>Pre-1920</td>
<td>Away from building</td>
<td>18.95</td>
<td><strong>&lt;0.01</strong></td>
</tr>
<tr>
<td>E</td>
<td>1920-70</td>
<td>Away from building</td>
<td>3.01</td>
<td>0.083</td>
</tr>
<tr>
<td>F</td>
<td>Post-1970</td>
<td>Away from building</td>
<td>8.29</td>
<td>0.004**</td>
</tr>
</tbody>
</table>

Figure 6.18 Box-plots of soil Pb (Box-Cox) for each building age class in the Greater Glasgow G-BASE dataset. Summary statistics for each group are given in the table below, with ranks for median soil Pb values.

6.4.1.9. Building distance

Studies have found higher soil Pb levels closer to older buildings, due to more Pb contamination closer to the building. However, this relationship was not observed in the Glasgow G-BASE dataset, as proven by the low and non-significant Wald value (Table 6.6C) and wide scatter when plotting soil Pb (Box-Cox) versus building distance (log) (Figure 6.19).

This relationship may not occur in Glasgow for several reasons. Firstly, previous studies only found a relationship between soil Pb and building distance for wooden painted homes, which are not common in Glasgow. Secondly, the soil Pb ~ building distance relationship is strongest in older houses and this analysis determined the
distance to any building. Thirdly, soil Pb elevation around a building only occurs at the very fine scale i.e. <5 m (Schwarz et al., 2012). Therefore, as the building age polygons included gardens, and the G-BASE sample is composed of samples from a 20 x 20 m area, the fine scale influence of buildings may be masked. Lastly, multiple 0 m (log = -4.2) distances with a range of soil Pb concentrations are skewing the results.

6.4.1.10. Organic matter

Organic matter has been shown to have a strong positive correlation with soil Pb in numerous studies (e.g. Gangloff et al., 2014; Chotpantarat et al., 2015). This is because negative functional groups or ligands in OM strongly bind positively charged Pb via chelation. Therefore, Pb’s mobility is significantly reduced in organic soils, thus more Pb is retained for longer.

This was shown to be the case in the Glasgow G-BASE dataset, with soil OM (log) having the strongest relationship with soil Pb of any covariate, as this model generates the highest Wald value (507 p<<0.01; Table 6.6C). The scatterplot confirms the strong relationship and shows that it is positive (Figure 6.20A). This relationship is likely strengthened as the Glasgow study area has a wide range of soil OM (2-36%, Table 6.6A) due to varying geology and topography, i.e. organic rich peats in the uplands, compared with more mineral soils above sandy alluvial or glacial till deposits towards the River Clyde (Fordyce et al., 2012).
Figure 6.20: Scatterplot comparing (A) organic matter (log) and (B) pH with soil Pb (Box-Cox) for Glasgow G-BASE samples.

6.4.1.11. pH

The pH in the Glasgow G-BASE dataset ranges from acidic (3.39) to alkali (8.79) with a mean of 5.3 (median=5.2; Table 6.6A). Two sites with a pH 0 were removed before analysis due to measurement error. Acidic pH values in Glasgow are associated with the presence of organic matter in the topsoil, which lowers the soil pH via humic and organic acids. However, anthropogenic additions such as industrial waste, coal waste, and calcareous building rubble have raised pH values across much of the city (Fordyce et al., 2012).

In the Glasgow G-BASE dataset pH and soil Pb do not have a significant relationship as the Wald value is low (1.21) and not significant ($p=0.27$). The scatterplot shows a very weak positive trend between pH and soil Pb (Figure 6.20B). This is because pH can influence Pb retention, but to a much lesser extent than OM. Specifically, in acidic soils Pb desorbs from complexes and Pb minerals dissolve (Stille et al., 2011). However, organic soils are often acidic due to the presence of humic acids, which may be complicating the pH~Pb relationship. In soils with a higher pH (>5) there are fewer H+ ions to compete with Pb$^{2+}$ at binding sites, so more Pb is retained (Vega et al., 2010). These processes cause pH to exert some control on total Pb levels but pH is more important in controlling the bioavailable, rather than the total soil Pb (Lacarce et al., 2012).
6.4.1.12. Soil texture

The analysis for the present study indicates that soil texture and Pb do have a significant relationship (Wald = 13.39, p=0.0012). However, this may be another example of correlation but not causation, as the relationship does not follow the expected pattern. Specifically, coarser soils (i.e. sands) have higher Pb levels than finer soils (i.e. clays) (Figure 6.19). The literature suggests that finer textured soils tend to retain higher levels of Pb, as clay minerals provide binding sites for Pb (Walraven et al., 2013a). Furthermore, through-flow tends to be higher in coarser soils, resulting in more metal particles being washed away.

The unexpected pattern may be due to uneven group sizes, as the majority of soils in Glasgow are silty (85%), with very few samples from sandy or clayey soils (Figure 6.19). However, Glasgow city centre is situated on the Clyde River valley where there are mostly sandy alluvial soils. The city centre also has generally higher soil Pb levels due to more pollution occurring in this region.

6.4.1.13. Summary

Table 6.10 summarises the results from the analysis of the Greater Glasgow G-BASE dataset. With covariates highlighted in green possible candidates for inclusion in future predictive models.

![Box-plots of soil Pb (Box-Cox) for each soil texture class in the Greater Glasgow G-BASE dataset. Summary statistics for each group are given in the table to the right, with ranks for median soil Pb values.]

![Table 6.10: Summary of Wald and p values for each covariate, ranked by Wald value.]

<table>
<thead>
<tr>
<th>Wald</th>
<th>p</th>
<th>Covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>506.78</td>
<td>&lt;0.01**</td>
<td>OM (log)</td>
</tr>
<tr>
<td>41.59</td>
<td>&lt;0.01**</td>
<td>Building age</td>
</tr>
<tr>
<td>21.37</td>
<td>&lt;0.01**</td>
<td>Hist. Ind. density (log)</td>
</tr>
<tr>
<td>13.39</td>
<td>0.0012**</td>
<td>Soil texture</td>
</tr>
<tr>
<td>11.91</td>
<td>0.018</td>
<td>Historic industry age</td>
</tr>
<tr>
<td>10.8</td>
<td>0.055</td>
<td>Land use</td>
</tr>
<tr>
<td>6.84</td>
<td>0.032*</td>
<td>Road type</td>
</tr>
<tr>
<td>3.84</td>
<td>0.057</td>
<td>Historic industry type</td>
</tr>
<tr>
<td>2.28</td>
<td>0.131</td>
<td>Building distance (log)</td>
</tr>
<tr>
<td>1.21</td>
<td>0.27</td>
<td>pH</td>
</tr>
<tr>
<td>0.02</td>
<td>0.88</td>
<td>Historic ind. distance (log)</td>
</tr>
<tr>
<td>0.01</td>
<td>0.93</td>
<td>Road distance (log)</td>
</tr>
</tbody>
</table>
### 6.4.2. Neighbourhood scale results - Paisley and Bishopbriggs

Table 6.11: Summary of statistical results for each covariate in Paisley (P) and Bishopbriggs (BB), including (A) summary statistics for continuous variables and group options for categorical variables, (B) variogram parameters, (C) model validation statistics, and (D) statistics exploring the relationship with soil Pb. **>>**99% confidence. *>>95% confidence. Wald ranks by study area. Green shading = 'best' validation result. HL = historic industry. \(^1\) difference from no covariate AIC.

<table>
<thead>
<tr>
<th>Covar.</th>
<th>None</th>
<th>OM (log)</th>
<th>Land use</th>
<th>Building dist.</th>
<th>Road dist.</th>
<th>Building Age</th>
<th>HL. Dist.</th>
<th>HL. Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Area</td>
<td>P</td>
<td>BB</td>
<td>P</td>
<td>BB</td>
<td>P</td>
<td>BB</td>
<td>P</td>
<td>BB</td>
</tr>
<tr>
<td>A - Summary statistics/group options</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>14.26</td>
<td>6.5</td>
<td>Back garden</td>
<td>24.65</td>
<td>12.95</td>
<td>20.22</td>
<td>18.66</td>
<td></td>
</tr>
<tr>
<td>Median</td>
<td>13.35</td>
<td>6</td>
<td>Commercial</td>
<td>10.57</td>
<td>7.25</td>
<td>12.88</td>
<td>9.45</td>
<td></td>
</tr>
<tr>
<td>Minimum - Maximum</td>
<td>3.1 - 43.63</td>
<td>2.9 - 14.8</td>
<td>Front garden</td>
<td>1 - 122</td>
<td>1 - 62</td>
<td>1 - 102</td>
<td>2 - 74</td>
<td></td>
</tr>
<tr>
<td>Octile skew</td>
<td>0.22</td>
<td>0.4</td>
<td>Road verge</td>
<td>0.77</td>
<td>0.68</td>
<td>0.54</td>
<td>0.74</td>
<td></td>
</tr>
<tr>
<td>Woodland</td>
<td></td>
<td></td>
<td>Woodland</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>33.76</td>
<td>75.42</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0 - 136</td>
<td>0 - 212</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>B - Variogram parameters</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nugget</td>
<td>0.0005</td>
<td>0.002</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Partial sill</td>
<td>0.04</td>
<td>0.0016</td>
<td>0.015</td>
<td>0.0043</td>
<td>0.028</td>
<td>0.0041</td>
<td>0.032</td>
<td>0.0013</td>
</tr>
<tr>
<td>Eff. Range</td>
<td>71.61</td>
<td>115.82</td>
<td>73.83</td>
<td>142.9</td>
<td>67.54</td>
<td>52.58</td>
<td>63.18</td>
<td>113.61</td>
</tr>
<tr>
<td>Lambda</td>
<td>-0.29</td>
<td>-0.46</td>
<td>-0.34</td>
<td>-0.31</td>
<td>-0.3</td>
<td>-0.44</td>
<td>-0.3</td>
<td>-0.46</td>
</tr>
<tr>
<td>C - Model validation statistics</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AIC</td>
<td>1462</td>
<td>1787</td>
<td>1419</td>
<td>1721</td>
<td>1453</td>
<td>1783</td>
<td>1427</td>
<td>1755</td>
</tr>
<tr>
<td>AIC diff(^1)</td>
<td>n/a</td>
<td></td>
<td>43</td>
<td>66</td>
<td>9</td>
<td>4</td>
<td>35</td>
<td>32</td>
</tr>
<tr>
<td>Mean SSE.</td>
<td>0.96</td>
<td>1.17</td>
<td>1.09</td>
<td>1.08</td>
<td>1</td>
<td>1.2</td>
<td>1.01</td>
<td>1.14</td>
</tr>
<tr>
<td>Med SSE.</td>
<td>0.48</td>
<td>0.5</td>
<td>0.64</td>
<td>0.58</td>
<td>0.52</td>
<td>0.59</td>
<td>0.58</td>
<td>0.53</td>
</tr>
<tr>
<td>D - Relationship with soil Pb</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald value</td>
<td>n/a</td>
<td></td>
<td>57.35</td>
<td>91.73</td>
<td>24.21</td>
<td>19.43</td>
<td>10.07</td>
<td>5.91</td>
</tr>
<tr>
<td>(p)</td>
<td>&lt;0.01(^**)</td>
<td>&lt;0.01(^**)</td>
<td>&lt;0.01(^**)</td>
<td>&lt;0.01(^**)</td>
<td>0.016(^*)</td>
<td>&lt;0.01(^**)</td>
<td>&lt;0.01(^**)</td>
<td>0.93</td>
</tr>
</tbody>
</table>
6.4.2.1. Organic matter

In the Glasgow neighbourhood study areas, OM (log) has the strongest relationship with soil Pb, as it has the highest Wald values (i.e. 57 and 92; Table 6.11D). OM (log) and soil Pb have a positive relationship which is visible in the scatterplots (Figure 6.22). This relationship is present in both study areas which further validates the results. The OM values were log transformed due to a significant skew, however there is also a strong positive relationship between soil Pb and non-transformed OM values. These results agree with analysis of the citywide G-BASE dataset, where a strong positive relationship between soil Pb and OM (log) was also clear (Table 6.6C; Figure 6.20A).

Figure 6.22: Scatterplots showing the relationship between modelled soil Pb (Box-Cox) and OM (log), in (A) Paisley and (B) Bishopbriggs, with best-fit models and adjusted $R^2$ values.
6.4.2.2. Land use

After OM, land use has the next strongest relationship with soil Pb both in Paisley and Bishopbriggs as proven by p values < 0.01, and high Wald values (>19; Table 6.11D). This relationship is slightly stronger in Paisley than Bishopbriggs, as indicated by a lower p value (by a factor of 10), and a higher Wald value (increase of five; Table 6.11D). This could occur because land use is more consistent over larger areas in Paisley compared with Bishopbriggs; therefore, nearby samples are more likely to belong to the same land use class in Paisley. Some general trends in soil Pb concentrations between land uses can be observed (Figure 6.23). This is discussed in Section 6.4.3.

![Box plots of Soil Pb concentrations by land use in Bishopbriggs and Paisley](image)

**Figure 6.23:** Box-plots for each land use class in (A) Bishopbriggs and (B) Paisley. Using modelled and Box-Cox transformed soil Pb values. Shaded by median back-transformed rank (darkest = highest). Summary statistics for each class are also provided.
6.4.2.3. Building distance

Building distance has the 4th strongest relationship with soil Pb in Paisley and Bishopbriggs. This is slightly stronger in Paisley (Wald = 10) than Bishopbriggs (Wald = 6), and $p$ is significant at the 99% confidence interval in Paisley but not Bishopbriggs ($p = 0.016$; Table 6.11D).

In Paisley the relationship is negative, with higher soil Pb levels closer to buildings (Figure 6.24A). This agrees with the literature, as Pb sources from buildings may contaminate nearby soil. Conversely, the relationship is positive in Bishopbriggs, as soil Pb levels increase away from buildings (Figure 6.24B). This is likely due to industrial contamination in a park occurring away from buildings, which will skew the underlying relationship (see Section 6.4.3).

![Figure 6.24: Scatterplots showing the relationship between modelled soil Pb (Box-Cox) and building distance (m), in (A) Paisley and (B) Bishopbriggs. With best-fit linear models and adjusted $R^2$ values.](image-url)
6.4.2.4. Road distance

Road distance was ranked 3rd in Bishopbriggs and 4th in Paisley, with regards to its relationship with soil Pb (Table 6.11D). In both study areas the Wald value is relatively high (>7) and significant, indicating that soil Pb does vary with road distance, but the relationship is not strong. However, when the values are plotted there is a wide scatter, and the best fit line is surprisingly positive in Bishopbriggs (Figure 6.25B). Other studies have found a negative relationship between soil Pb and road distance, with higher soil Pb levels immediately next to roads due to historic leaded-petrol emissions (see references in Table 6.3). This was observed in Paisley (Figure 2.25A) but not Bishopbriggs. The negative relationship in Bishopbriggs is likely due to elevated soil Pb levels next to an old refuse heap in the middle of a park (see Section 6.4.3). Therefore, these samples are located further from roads and have high soil Pb contents, so are likely impacting the soil Pb and road distance relationship.

![Graph A - Paisley](image)

**Figure 6.25:** Scatterplots showing the relationship between modelled soil Pb (Box-Cox) and road distance (m), in (A) Paisley and (B) Bishopbriggs. With best-fit linear models and adjusted $R^2$ values.
6.4.2.5. Building age

Building age has a relatively strong relationship with soil Pb levels in Paisley (ranked 3rd most influential), but not in Bishopbriggs. This is indicated by the high, and significant, Wald value in Paisley (14.2), suggesting that there is a significant difference in soil Pb levels between the building age classes. This is not the case in Bishopbriggs, as the Wald value is very low (1.37) and not significant (Table 6.11D).

There is no statistical difference in Pb levels between soils adjacent to buildings and soils away from buildings (Wald = 2.61 and 0.34, \( p = 0.11 \) and 0.56 in Paisley and Bishopbriggs respectively). However, in Paisley soil away from buildings has the 2nd lowest median level (Figure 6.26), so there is an observable difference. Conversely, in Bishopbriggs soil away from buildings has the 2nd highest median Pb level because it includes contaminated soil in Etive Crescent park.

![Box-plots for each building age class in (A) Paisley and (B) Bishopbriggs. Using modelled and Box-Cox transformed soil Pb values. Shaded by median back-transformed rank. Summary statistics for each class are also provided.](image)

<table>
<thead>
<tr>
<th>Building age</th>
<th>None</th>
<th>1896</th>
<th>1930</th>
<th>1950</th>
<th>1992</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>43</td>
<td>4</td>
<td>35</td>
<td>6</td>
<td>24</td>
</tr>
<tr>
<td>Median</td>
<td>127.37</td>
<td>247.42</td>
<td>200.92</td>
<td>80.00</td>
<td>236.91</td>
</tr>
<tr>
<td>Rank</td>
<td>4</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Building age</td>
<td>None</td>
<td>1896</td>
<td>1930</td>
<td>1950</td>
<td>1970</td>
</tr>
<tr>
<td>B</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N</td>
<td>44</td>
<td>8</td>
<td>29</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Median</td>
<td>198.66</td>
<td>178.93</td>
<td>186.66</td>
<td>205.54</td>
<td>183.95</td>
</tr>
<tr>
<td>Rank</td>
<td>2</td>
<td>5</td>
<td>3</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 6.26: Box-plots for each building age class in (A) Paisley and (B) Bishopbriggs. Using modelled and Box-Cox transformed soil Pb values. Shaded by median back-transformed rank. Summary statistics for each class are also provided.
6.4.2.6. Historic industry type and distance

Overall, there is no relationship between soil Pb and historic industry type in Paisley and Bishopbriggs. This is proven by the low (<3) and non-significant ($p > 0.5$) Wald values (Table 6.11D). The relationship is slightly stronger in Bishopbriggs, compared with Paisley. This is largely driven by a historic refuse heap in the northeast of Bishopbriggs, which likely contained Pb contaminated waste in the past. The median soil Pb level in the refuse heap is 256 mg kg$^{-1}$ (Figure 6.27).

![A - Paisley](image)

![B - Bishopbriggs](image)

<table>
<thead>
<tr>
<th>Historic Ind. type</th>
<th>Coal mining</th>
<th>Old quarry</th>
<th>Railway land</th>
<th>Steel works</th>
<th>Non-industrial</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>13</td>
<td>6</td>
<td>2</td>
<td>9</td>
<td>108</td>
</tr>
<tr>
<td>Median</td>
<td>222.59</td>
<td>148.86</td>
<td>148.43</td>
<td>255.78</td>
<td>175.16</td>
</tr>
<tr>
<td>Rank</td>
<td>2</td>
<td>4</td>
<td>5</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

**Figure 6.27:** Box-plots for each historic industry type class in (A) Paisley and (B) Bishopbriggs. Using modelled and Box-Cox transformed soil Pb values. Shaded by median back-transformed rank. Summary statistics for each class are also provided.
In Paisley it is surprising that soil surrounding the steel works does not have elevated Pb levels, as this type of industry is likely to have generated Pb pollution (Table 6.5). The median Pb level in these soils is unexpectedly low (81 mg kg$^{-1}$), which may be linked to low organic matter levels in these soils or simply that the steel production process used at Ferguslie plant did not emit much Pb.

There was also no significant difference in Pb levels between industrial and non-industrial soil (Wald = 1.35 and 0.05, in Paisley and Bishopbriggs respectively). In fact non-industrial soil has the 2nd highest median Pb level in Paisley (Figure 6.27). This is likely because not all types of industry generated Pb pollution, and other factors are often more important in controlling Pb levels e.g. organic matter. Some types of historic industry are very old, for example the old quarries date from before 1869, but likely generated minimal Pb pollution.

There is also no relationship between soil Pb and distance to historic industry in Paisley (Wald = 0), and only an incredibly weak relationship in Bishopbriggs (Wald = 3.27; $p = 0.07$; Table 6.12D). In Bishopbriggs the relationship is negative (Figure 6.28B), and again is largely driven by the refuse heap in Etive Crescent park, where soil Pb concentrations can been seen to decrease away from the site. However, as not all types of industry generated Pb pollution, as there is not a consistent pattern of decreasing soil Pb away from all historic industry polygons.

**Figure 6.28:** Scatterplots showing the relationship between modelled soil Pb (Box-Cox) and historic industry distance (m), in (A) Paisley and (B) Bishopbriggs. With best-fit linear models and adjusted $R^2$ values.
6.4.2.7. Summary

Table 6.12 summarises the results from the neighbourhood analysis and provides select Greater Glasgow results for comparison.

Table 6.12: The relationship between soil Pb and each covariate, ranked (R) by Wald value from most to least influential. With results from the neighbourhood studies and the Glasgow G-BASE analysis for comparison. Note – five additional covariates were included in the Greater Glasgow analysis but are not shown here (see Table 6.10).

<table>
<thead>
<tr>
<th>Area</th>
<th>Covar.</th>
<th>Paisley</th>
<th>Bishopbriggs</th>
<th>Greater Glasgow</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Wald</td>
<td>p</td>
<td>R</td>
<td>Wald</td>
</tr>
<tr>
<td>OM (log)</td>
<td>57.4</td>
<td>&lt;0.01**</td>
<td>1</td>
<td>91.7</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>506.8</td>
</tr>
<tr>
<td>Land use</td>
<td>24.2</td>
<td>&lt;0.01**</td>
<td>2</td>
<td>19.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>10.8</td>
</tr>
<tr>
<td>Road dist.</td>
<td>7.7</td>
<td>&lt;0.01**</td>
<td>5</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.01</td>
</tr>
<tr>
<td>Build dist.</td>
<td>10.1</td>
<td>&lt;0.01**</td>
<td>4</td>
<td>5.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.28</td>
</tr>
<tr>
<td>Build age</td>
<td>14.2</td>
<td>&lt;0.01**</td>
<td>3</td>
<td>1.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>41.69</td>
</tr>
<tr>
<td>HI dist.</td>
<td>0</td>
<td>0.97</td>
<td>7</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.02</td>
</tr>
<tr>
<td>HI type</td>
<td>1.4</td>
<td>0.71</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>3.8</td>
</tr>
</tbody>
</table>

6.4.3. The importance of location

The importance of within-city location in influencing the covariate and soil Pb relationship was examined by comparing results from Paisley and Bishopbriggs (Table 6.12). The difference in covariates influence between cities is discussed in Chapter 9.

In both Paisley and Bishopbriggs OM is clearly the most influential covariate (Table 6.12). This is despite the larger range and higher average OM levels in Paisley (range = 3-44%, mean = 14%), compared with Bishopbriggs (range = 3-15%, mean = 7%; Table 6.12A). The importance of OM in both neighbourhoods, and the Greater Glasgow G-BASE dataset, suggests that OM is a dominant control on soil Pb throughout Glasgow. This was also observed in other urban studies, especially those focusing on bioavailable Pb (e.g. Lu et al., 2011; Attanayake et al., 2014; Wragg et al., 2019; Shi et al., 2021). As discussed in Section 6.1.7, Pb tends to have a lower mobility in organic soils because OM strongly binds Pb (Gangloff et al., 2014). However, OM only retains Pb deposited to the soil, it is not a direct source of Pb. Thus, the relationship between soil Pb and OM is dependent on background
atmospheric Pb deposition and point sources of Pb e.g. petrol, paint, and industry. These sources of Pb are/were relatively consistent across Glasgow, as shown by similar median soil Pb concentrations in Paisley (151 mg kg\(^{-1}\)), Bishopbriggs (170 mg kg\(^{-1}\)), and Greater Glasgow (135 mg kg\(^{-1}\); Table 3.2).

Land use also significantly influences soil Pb levels, regardless of location in Glasgow (Table 6.12). Generally, private land uses i.e. commercial and residential, are most at risk of having high soil Pb levels (Figure 6.23). Soils in these land uses likely accumulated Pb from multiple sources, such as old Pb-paint, historic leaded-petrol emissions, background historical industrial emissions, and dumping of coal ash from fireplaces (see references in Section 6.1.1.). Conversely, public land uses i.e. recreational land, parks, road verges, and urban open space, tend to have lower soil Pb levels (Figure 6.23) due to fewer Pb sources at a greater distance.

There are some anomalies with regards to land use and soil Pb in the neighbourhood datasets. For example, there is a large disagreement between soil Pb levels in open space and residential open space, with each ranked the alternating highest and lowest in Bishopbriggs and Paisley (Figure 6.23). This is likely caused by former land use. For instance, in Bishopbriggs an old refuse heap in Etive Crescent Park, probably containing Pb-contaminated material, occurs in open space land (Figure 6.29). In Paisley previously contaminated topsoil from historic railway lines may have been reworked during new building next to residential open space (Figure 6.30). This is also shown in the unexpectedly high median soil Pb level (262 mg kg\(^{-1}\)) for front gardens in Paisley. These disparities between land uses highlights the importance of local conditions, when examining the relationship between soil Pb and a covariate.

Road and building distance have a significant correlation with soil Pb in both neighbourhood datasets (Table 6.12). However, these relationships are unexpectedly positive in Bishopbriggs (Figures 6.24B and 6.25B), possibly due to contaminated soil in Etive Crescent Park occurring far from roads and buildings (Figure 6.29). This is evident in the scatterplots (Figures 6.24B and 6.25B) as a few high soil Pb outliers skewing the relationship. These outlying points could have been removed and the analysis rerun.
Figure 6.29: Map of Etive Crescent refuse heap in Bishopbriggs with soil Pb concentrations at each sample point, historic industry locations, and building age.

Figure 6.30: Map of historic railways in Paisley with soil Pb concentrations at each sample point, historic industry locations, and building age.
Building age does not have a consistent relationship with soil Pb in either Paisley and Bishopbriggs (Table 6.12). In Paisley the relationship generally follows the expected pattern, with soil around older buildings exhibiting higher Pb levels (Figure 6.26A), due to aforementioned Pb sources (see Section 6.1.5.). One exception is that soil around 1992 buildings have the 2nd highest median Pb level (237 mg kg\(^{-1}\); Figure 6.26A). This is likely due to historic land use, as the 1992 houses are located at the site of the old industrial railways (Figure 6.30). In Bishopbriggs the story is very different, as soil around newer, 1950s buildings have the highest median Pb levels (206 mg kg\(^{-1}\)), and soil around older, pre-1896 buildings have the 5th highest median (179 mg kg\(^{-1}\); Figure 6.26B). Lead levels are likely elevated in soil adjacent to newer buildings as these are located next to Etive Crescent Park, which contains a historic refuse heap (Figure 6.29). The relationship between soil Pb and building age in Paisley but absence of one in Bishopbriggs may also arise because Paisley has a simpler building age history, with more larger areas of buildings erected in the same period. Whereas, building in Bishopbriggs occurred in smaller scattered areas and other factors e.g. Pb from waste, may have overprinted any relationship.

Historic industry type and distance do not correlate with soil Pb in either Paisley or Bishopbriggs. In Bishopbriggs a refuse heap is clearly contaminating nearby soil with Pb, but another refuse heap in the south of the study area does not show elevated soil Pb concentrations. This suggests that soil Pb levels cannot be inferred for one type of industry, as they may differ from site to site. Other types of past industries e.g. railways and old quarries, have similar Pb concentrations in Paisley and Bishopbriggs (difference < 30 mg kg\(^{-1}\), Figure 6.27), so may be less variable with regards to Pb content. However, soils around most historic industrial sites are not elevated, when compared with non-industrial soils (Figure 6.27).

Overall, local factors can play an important role in the relationship between soil Pb and covariates in each neighbourhood. For example, contamination from an old refuse heap in Bishopbriggs, and historic railway lines in Paisley, can skew the expected relationships between soil Pb and some covariates. These local disparities will make it difficult to predict soil Pb concentrations using covariates at the neighbourhood scale. Even if the neighbourhoods are located in the same city.
6.4.4. The importance of scale

The importance of scale with regards to soil Pb’s relationship with each covariate, was explored by comparing results from Glasgow e.g. neighbourhood versus city-wide results (Table 6.12). Some covariates have a similar relationship with soil Pb regardless of scale, whereas others differ greatly.

Firstly, OM has a very consistent relationship with soil Pb at both the neighbourhood and city scale (Table 6.12). As discussed previously (Section 6.4.3.) the relationship between soil organic matter and Pb is dependent on the amount of Pb deposited to the soil. Therefore, if the background atmospheric Pb levels were relatively consistent across Glasgow, then the relationship between soil Pb and OM also should remain the same.

Road and building distance are covariates that are clearly influenced by scale. Specifically, road and building distance and soil Pb have a high and significant Wald value in Paisley (7.7 and 10.1 respectively, \( p << 0.01 \); Table 6.12), but do not in the Glasgow-wide dataset (Wald = 0.01 and 2.28, \( p > 0.1 \); Table 6.10). The Bishopbriggs results should be discounted as they are skewed by the refuse heap contamination. The strong relationship between soil Pb and road or building distance in Paisley is likely because of the more detailed sampling interval (42 m vs 500 m) and smaller sample aggregation area (1 m vs 20 m), compared with the G-BASE samples (see Chapter 3 for details). However, the G-BASE dataset aimed to record soil Pb at the city scale, which meant detailed sampling was not possible. This highlights the importance of sampling scale when investigating the relationship between soil Pb and a covariate.

Building age has a stronger relationship with soil Pb at the city scale (Wald = 42), compared with the neighbourhood scale, especially in Bishopbriggs (Wald = 1; Table 6.12). However, each scale uses different building age classes, making direct comparison difficult. For instance, only four building age classes were digitised at the city scale (due to the availability of city-wide maps), with an average of 321 samples per group (Figure 6.18). At the neighbourhood scale there are five to six building age classes, with an average of 23 samples per group (Figure 6.26). This 14-fold increase in the number of samples in each building age class, means the relationship with soil Pb will be less influenced by outliers at the city scale. For example, in Bishopbriggs the refuse heap at Etive Crescent park increases soil Pb...
concentrations in several samples collected from nearby 1950s houses. These outlier samples would have been less influential in the larger city-wide dataset.

Conversely, land use has a slightly weaker relationship with soil Pb at the city scale (Wald = 11), compared with the neighbourhood scale (Wald = 19 and 24; Table 6.12). The ranking of land use with regards to soil Pb is also broadly similar in all study areas in Glasgow. For example, individual Wald tests demonstrated that private land uses have significantly higher soil Pb concentrations than public land uses. In the neighbourhood study areas residential/non-residential open space have higher than expected soil Pb concentrations due to past land use e.g. the refuse heap in Etive Crescent Park in Bishopbriggs, and reworking of historic railway soil in Paisley (Figures 6.29 and 6.30). As with building age, these outliers are diluted in the more extensive Glasgow wide dataset.

Generally, the sampling scale has a mixed impact on the relationship between soil Pb and the seven covariates in Table 6.12. It depends on the covariate in question. For some covariates, the more spatially detailed sampling highlights their underlying relationship with soil Pb. For example, the coarser and aggregate sampling stagey in the G-BASE study likely masked any relationship present between soil Pb and road distance. In other cases, the smaller scale sampling obscures the relationship between soil Pb and certain covariates. For instance, there is no relationship between soil Pb and building age in Bishopbriggs due to anomalous examples skewing the results. The city-wide analysis generalises these relationships, consequently outliers have less impact on overall results. Ideally a mix of sampling scales i.e. nested sampling, would be required to accuracy account for the influence of covariates on soil Pb in a complex urban environment. Otherwise, separate models should be generated for different scales.

6.4.5. Modelling and mapping soil Pb

Including covariates in soil Pb models and maps can often explain Pb variation and consequently lead to less uncertain mapped predictions. This was explored in the Paisley and Bishopbriggs datasets using semivariogram parameters and cross-validation statistics.

In Paisley, the OM, building age, and then land use, variograms show the biggest visual difference from the no covariates variogram (Figure 6.31). With the road distance, and historic industry distance and type models being very similar to the
no covariate model (Figure 6.31). This is backed up by variogram parameters i.e. the partial sill is much lower in the OM, building age and land use models, compared with the no covariate model (Table 6.11B), suggesting these covariates are accounting for variance in soil Pb between nearby points. Although, they have slightly different lambdas so are not directly comparable. Similarly, in Bishopbriggs the OM, land use and road distance models vary most from the no covariate model, but they have different smoothing parameters, making it very tricky to directly compare (Figure 6.32).

The OM (log) models have the lowest AIC in both areas, which are substantially lower than the no covariate AICs (Table 6.11D), suggesting adding OM to the model significantly improves it. In Paisley all the covariates have an AIC lower than the no covariate model apart from historic industry and distance, and only building age produces a higher AIC in Bishopbriggs. This corresponds with previous analysis showing that these covariates have no correlation with soil Pb in the study area in question.

If you examine the 10-fold cross validation results (Table 6.11D) then the ‘best’ model at predicting soil Pb appears to be the no covariate model (median SSE closest to 0.45) or the land use model (mean SSE closest to 1) in Paisley. In Bishopbriggs the models with the best validation statistics are the historic industry distance model (median SSE closest to 0.45) or the OM log model (mean SSE closest to 1).

The E-BLUP maps (Figures 6.33 and 6.34) show little difference from the simply kriged map i.e. no covariates, apart from with land use as a covariate. In Paisley and Bishopbriggs, the impact of the lower mean soil Pb value for road verges is clear as linear features bisecting hotspots. Building age and road distance have subtle alterations when compared with the no covariate map in both areas. There are also slight differences in the historic industry distance/type maps in Bishopbriggs, but not clear changes in Paisley.

Overall, there may not be a single ‘best’ model or covariate when mapping soil Pb. Additionally, a combination of multiple covariates should be explored. This will be investigated in more detail when using machine learning (Chapter 8).
Figure 6.31: Variograms modelling soil Pb in Paisley, with different covariates. The no covariate models are shown on each plot for comparison, along with method of moment (MOM) points to assess model fit. The same scales are used in all plots, but lambda values used in each model do vary (see Table 6.11B).
Figure 6.32: Variograms modelling soil Pb in Bishopbriggs, with different covariates. The no covariate models are shown on each plot for comparison, along with method of moment (MOM) points to assess model fit. The same scales are used in all plots apart from OM log, and lambda values used in each model do vary (see Table 6.11B).
Figure 6.33: Soil Pb maps (E-BLUP) of Paisley with different covariates used during kriging. The ordinary kriged soil Pb map i.e. no covariates, is displayed as map A for comparison. Classification is based on percentiles of the simply kriged map i.e. <5th, 5th-25th, 25th-50th, 50th-75th, 75th-95th, and >95th percentile, averaged over Paisley and Bishopbriggs.
**Figure 6.34:** E-BLUP soil Pb maps of Bishopbriggs with different covariates used during kriging. The ordinary kriged soil Pb map i.e. no covariates, is displayed as map A for comparison. Classification is based on percentiles of the simply kriged map i.e. <5\(^{th}\), 5\(^{th}-25\(^{th}\), 25\(^{th}-50\(^{th}\), 50\(^{th}-75\(^{th}\), 75\(^{th}-95\(^{th}\), and >95\(^{th}\) percentile, averaged over Paisley and Bishopbriggs.
6.5. **Conclusions**

Overall, there are some general similarities in the relationship between each covariate and soil Pb in the different study areas, but there are also important disagreements. The first aim of this chapter was to rank covariate importance in each study area, in order to decide which covariates to include in future machine learning methods (Chapter 8). Secondly, this analysis aimed to determine if these covariates have the same influence on soil Pb concentrations regardless of location and scale? This will help explore whether the machine learning models are applicable to other locations and scales.

At the Glasgow city scale only the covariates with a significant (or almost significant) Wald value i.e. \( p<0.06 \), should be included in future models. Additionally, their correlation with soil Pb should generally agree with the literature. Therefore, although soil texture and road type have significant Wald values, they will not be included in future models as they do not follow the expected pattern with regards to soil Pb. Overall, this chapter shows that OM, building age, historic industry density, age, and type, and land use should be included in future predictive models of soil Pb in Greater Glasgow.

At the neighbourhood scale OM (log) is the most influential covariate in Paisley and Bishopbriggs, then land use, road and building distance, and building age (only in Paisley). With historic industry distance and type having very little influence on soil Pb levels, especially in Paisley. Therefore, in Paisley only the five covariates that showed a significant relationship with soil Pb should be included in future predictive models. In order of strongest relationship to weakest, based on overall Wald values, these are OM (log), land use, building age, building distance, and road distance. In Bishopbriggs five covariates have a significant relationship with soil Pb, namely OM (log), land use, road and building distance, and historic industry type. However, road and building distance have unexpectedly positive relationships with soil Pb due to high Pb samples in the middle of Etive Crescent Park. Therefore, they will not be included in future predictive models to prevent over-fitting.

Location appears to have a strong control on the importance of some covariates at the neighbourhood scale. There are some similarities between the neighbourhoods studied here e.g. OM and land use are important predictors in
both Paisley and Bishopbriggs; however, there are also major differences caused by local factors. For example, in Paisley reworking of contaminated railway soil during a new housing development has unexpectedly increased Pb concentrations in some soil. Similarly, in Bishopbriggs an old refuse heap likely containing Pb material has elevated nearby soil Pb values but also skewed the relationship between soil Pb and some covariates i.e. road and building distance. This suggests that neighbourhood machine learning models may have limited transferability due to important differences in covariate influence between neighbourhoods. This is tested in Chapter 9.

Scale also impacts covariate influence but is dependent on the individual covariate. For example, road and building distance are more influential in the neighbourhood datasets, compared with the city-wide dataset, due to difference in sampling strategies. However, building age and historic industry benefit from the more numerous examples in the city dataset, which down weights the influence of any outliers. Ideally, nested sampling would be required to accurately account for the influence of covariates on soil Pb in a complex urban environment.

Lastly, this chapter showed how covariates can be used to improve kriging predictions of soil Pb concentrations at the small scale. This was achieved using the E-BLUP, with some covariates i.e. OM, substantially reducing the no covariate model’s AIC. Alternatively, covariates which have no relationship with soil Pb do not improve the maps e.g. building age in Bishopbriggs and historic industry type and distance in Paisley.
References - Chapter 6


Appendix 6.1: OS greenspace land use classification

Created via OS MasterMap topographic databases and aerial imagery

- Covers urban areas >6km² in England/Wales and areas with populations >500 in Scotland. No coverage in Northern Ireland. Updated every six months.
- During the spatial join samples did not always fall exactly on a greenspace polygon, so the closest polygon was used: average distance to land use polygon = 1 m, only 316/1285 samples >0 m from a polygon, max distance = 81 m.

<table>
<thead>
<tr>
<th>OS Greenspace primary function</th>
<th>Land use (this study)</th>
<th>Justification</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tennis court</td>
<td>Commercial and municipal</td>
<td>All classed as recreational land, but in G-BASE dataset these are not significantly different (Contrast I: Wald&lt;1) from other commercial/municipal land so grouped together.</td>
</tr>
<tr>
<td>Golf courses</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bowling greens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other sports</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Civic space</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Institutional grounds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cemetery/Churchyard</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Religious grounds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other functional greenspace</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Play space</td>
<td>Commercial and municipal</td>
<td>All fall under the municipal category. But when the G-BASE dataset was analysed, these had very similar soil Pb levels to recreational and school land, so all grouped together.</td>
</tr>
<tr>
<td>School grounds</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amenity - Residential or business</td>
<td>Residential open space</td>
<td>Matches the location of residential open space, used as a land use class in Paisley and Bishopbriggs. Also includes current industrial land.</td>
</tr>
<tr>
<td>Public Park or garden</td>
<td>Open space</td>
<td></td>
</tr>
<tr>
<td>Playing field</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Private gardens</td>
<td>Garden</td>
<td></td>
</tr>
<tr>
<td>Allotment/Community Gardens</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Woodland</td>
<td>Natural</td>
<td>All semi-natural land uses - could have separated woodland as own class, but no difference in soil Pb levels from other semi-natural classes.</td>
</tr>
<tr>
<td>Green access routes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Riparian routes</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public park and garden</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open water</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Open semi-natural</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amenity - transport</td>
<td>Road verge</td>
<td>Sites will be close to a road or railway line.</td>
</tr>
</tbody>
</table>
Appendix 6.2: Example Matlab code for geostatistical analysis

% Select variable and covariates
z = meanPb; %use mean soil Pb measurements as variable
OMlog = log(OM); %Transform OM

% Set up the covariance matrix
n = length(z);
M = ones(n,2);
M(:,2) = OMlog; %use OM (log) as the covariate
M = dummyvar(BA2); %alt. categorical covariate matrix (build age)

% Rescale coordinates to make computation simpler
X = [easting northing];
lx = max(X(:,1)) - min(X(:,1)); %range of eastings
mx = [min(X(:,1)) min(X(:,2))]; %minimum easting and northing
X(:,1) = (X(:,1) - mx(1))/lx; %rescale eastings
X(:,2) = (X(:,2) - mx(2))/lx; %rescale northings

% Make sure only have unique sites
k = find(floor(sample)~=sample);
X(k,1)=X(k,1)+0.000001;
[dum,a,b] = unique(sample);
X = X(a,:);
z = z(a);
M = M(a,:);
H = distance(X',X'); % Create a distance matrix

% REML fitting of variogram
pfixed = NaN*ones(6,1); %create matrix for 6 parameters: p1=nugget,
p2=partial sill, p3=range, p4=smoothness (limited to 1.5),
p5=lambda/transform, p6=measurement error
pfixed(6) = 0; %no measurement error as complicates results
options = []; %no extra options
% Variogram optimization
[L,par,C,zt,Lm,NI,beta] = ML_fit_opts(pfixed,H,z,M,options);
er = eff_range(par,max(max(H)))); % Calculate effective range
er = er*lx;
npar = length(beta)+(sum(isnan(pfixed)))); % Calculate AIC
AIC = 2*L+2*npar;

% Overall Wald test
L_i = [0 1];
Wald_i = (L_i*beta)'*inv(L_i*NI*L_i')*(L_i*beta);
P_i = 1-chi2cdf(Wald_i,1);

% Output the parameters
out_table = table(par(1), par(2), er, par(4), par(5), par(6), L,
AIC, Wald_i, P_i)

% Plot the results in scatterplot
scatter(OMlog, zt,'red', 'filled')
hold 'on'
\[ r_{OM} = \max(OM\log) - \min(OM\log) \]
\[ x = []; \]
\[ x(:,1) = \min(OM\log) - r_{OM} / 100; r_{OM} / 1000: r_{OM} / 100 + \max(OM\log); \]
\[ y = \beta(1) + \beta(2) \cdot x; \]
\[ \text{plot}(x,y,'k'); \]
\[ \text{title('Bishopbriggs')}; \]
\[ \text{xlabel('OM (log)');} \]
\[ \text{ylabel('Soil Pb (box-cox)');} \]

% If a categorical variable plot the results in boxplots for simulated results in original units
% Krige land use at new location of only 1 point in each land use
\[ T3 = \text{rand}(nq,2)+1000; \% \text{random coordinates away from known points} \]
\[ MT3 = \text{eye}(nq); \% \text{different building age at each point} \]
\[ [\text{out3, map\_mean3, map\_sd3, sim\_out3}] = \text{predict\_options}(\text{par}, X, T3, M, MT3, z, \text{options}); \]
% box plot using 1000 simulations
\[ zm = []; \]
\[ \text{Buildagem} = []; \]
\[ \text{uL} = \text{unique(BA2)}; \]
\[ \text{for} \ i = 1:nq; \]
\[ zm = [zm; \text{sim\_out3}(i,:)']; \]
\[ \text{Buildagem} = [\text{Buildagem}; \text{repmat(uL(i),1000,1)]}; \]
\[ \text{end} \]
\[ \text{subplot}(3,1,3) \]
\[ \text{boxplot}(zm,\text{Buildagem}); \]
\[ \text{xlabel('Building age')}; \]
\[ \text{ylabel('Soil Pb (mg/kg)');} \]

% 10-fold cross validation, WARNING takes a while to run
\[ \text{nfold}=10; \% \text{randomly allocate each measurement one of 10 folds} \]
\[ \text{fold}=\text{ceil}(\text{nfold} \cdot \text{rand}(\text{length}(z),1)); \]
% Rerun model without one fold and compare predicted and actual results, repeat 10 times
\[ \text{z\_cv} = \text{NaN*ones(length}(z),1); \]
\[ \text{zt\_cv} = \text{NaN*ones(length}(z),2); \]
\[ \text{v\_cv} = \text{NaN*ones(length}(z),1); \]
\[ \text{for} \ i = 1:\text{nfold}; \]
\[ f = \text{find}(\text{fold}==i); \]
\[ \text{nf} = \text{find}(\text{fold}~=i); \]
\[ [\text{Li, pari, Ci, zti, Lmi, WIi, beta1}] = \text{ML\_fit\_opts}(\text{pfixed}, \text{H}(\text{nf, nf}), z(\text{nf}), \text{M(\text{nf,:}, options}); \]
\[ [\text{outi,}, \text{sim\_outi}] = \text{predict\_options}(\text{pari}, X(\text{nf,:}), X(f,:), M(\text{nf,:}), \text{M(f,:), z(nf), options}); \]
\[ \text{zt\_cv}(f,1)=\text{outi(:,1)}; \]
\[ \text{zt\_cv}(f,2)=\text{bc}(z(f), pari(5)); \]
\[ \text{v\_cv}(f)=\text{outi(:,2)}; \]
\[ \text{z\_cv}(f)=\text{mean}(\text{sim\_outi}, 2); \]
\[ \text{end} \]
\[ \text{msse} = \text{mean}((\text{zt\_cv(:,1)-zt\_cv(:,2)}).\text{^2})./(\text{v\_cv}); \]
\[ \% \text{mean squared standardised prediction error should be 1} \]
\[ \text{medsse} = \text{median}((\text{zt\_cv(:,1)-zt\_cv(:,2)}).\text{^2})./(\text{v\_cv}); \]
\[ \% \text{mean standardised square error should be 0.45} \]
Chapter 7. The origins of soil Pb in Paisley and Bishopbriggs

In the previous chapter several controls on soil Pb concentrations were found. However, some covariates did not correlate with soil Pb concentrations, despite evidence in other studies that a covariate should have an influence. This lack of relationship between soil Pb and some covariates may be due to overprinting from multiple different sources. Therefore, I performed isotopic analysis on half the soil samples collected from Bishopbriggs and Paisley. I then used the resulting isotopic ratios to apportion soil Pb into that originating from petrol sources and from coal sources. The spatial pattern of soil Pb originating from petrol or coal was mapped. Additionally, two covariates: road and historic industry distance, were investigated to determine what was controlling these patterns.

7.1. Background

Lead isotopes are a common tool for determining the predominant provenance(s) of Pb in urban soils (Mirlean et al., 2005; Farmer et al., 2011; Mackay et al., 2013; Rodríguez-Seijo et al., 2015; Bi et al., 2018; Liang et al., 2019). Lead has four stable isotopes; $^{204}\text{Pb}$ (1%) which is primordial, and $^{206}\text{Pb}$ (24%), $^{207}\text{Pb}$ (23%) and $^{208}\text{Pb}$ (52%) which are the end products of the uranium (i.e. $^{238}\text{U}$, half-life $4.5 \times 10^9$ years; and $^{235}\text{U}$, half-life $0.7 \times 10^9$ years), and thorium ($^{232}\text{Th}$, half-life $14.1 \times 10^9$ years) radioactive decay series, respectively (Ryan, 2014). Lead ores have unique Pb isotope ratios i.e. varying proportions of $^{206}\text{Pb}/^{207}\text{Pb}$, $^{208}\text{Pb}/^{207}\text{Pb}$ and $^{206}\text{Pb}/^{208}\text{Pb}$, due to differences in the U/Th ratio and age of the parent ore (Li et al., 2012). Therefore, any products created from an ore will have the same isotopic ratio as the ore (Walraven et al., 2013).

The source of soil Pb can be determined from Pb’s isotopic composition because historically Pb used in petrol and in industrial activities had different $^{206}\text{Pb}/^{207}\text{Pb}$ ratios. In the UK, Associated Octel produced leaded-petrol by combining Broken Hill ore deposits from Australia ($^{206}\text{Pb}/^{207}\text{Pb} = 1.04$), with British Columbian ore deposits ($^{206}\text{Pb}/^{207}\text{Pb} = 1.16$) in a 70:30 mixture (Sugden et al., 1993). Therefore, pre-1998 UK petrol emissions had a low $^{206}\text{Pb}/^{207}\text{Pb}$ ratio ranging from 1.06-1.12, with a mean of 1.082 (Figure 7.1; Sugden et al., 1993; Farmer et al., 1996). British industrial emissions generally used indigenous Pb ore deposits, which are
differentiable from petrol emissions as they have a higher $^{206}\text{Pb}/^{207}\text{Pb}$ ratio, with ore from the Leadhills-Wanlockhead in southwest Scotland having a $^{206}\text{Pb}/^{207}\text{Pb}$ ratio of 1.173±0.002 (Figure 7.1; Farmer et al., 2000). Furthermore, British coal also had a higher $^{206}\text{Pb}/^{207}\text{Pb}$ ratio than petrol. Farmer et al. (1999) determined Scottish coals had a mean $^{206}\text{Pb}/^{207}\text{Pb}$ ratio of 1.181±0.011 and a mean $^{208}\text{Pb}/^{207}\text{Pb}$ ratio of 2.478±0.03 (Figure 7.1).

Soil Pb from industrial sources likely originated from coal fuel emissions, waste, and from coal mining. Scottish coals contain up to 137 mg kg$^{-1}$ of Pb, with an average of 24 mg kg$^{-1}$ (Farmer et al., 1999). Coal contains Pb because during geological processes the element is adsorbed by organics which were later compressed to form hydrocarbons (Alloway, 1995). Heating coal volatises the Pb, which is predominantly emitted as fly ash (Pacyna, 1987). These ultra-fine particles can be transported long distances, up to 1,000 km (Pandey et al., 2012; Inoue et al., 2014). Furthermore, coarser bottom ash particles are left in the furnace, and concentrate Pb. Furnace ash was often deposited on soil to act as a natural fertiliser. Lastly, coal mine waste or gaunge has high sulphide levels, which often contains Pb (Chen et al., 2014), and the acidic conditions in the gaunge facilitates oxidation and leaching of Pb (Fang et al., 2015). Therefore, coal Pb could have entered the soil via fine particles created from domestic and industrial coal burning, or deposition of coal waste; including gaunge from mining, and domestic fire ash remains.

There are some issues with using Pb isotopes to determine its source. Firstly, Pb used in British paint did not originate from a single Pb-ore; therefore, paint’s isotopic signature is harder to separate from petrol emissions. The average

---

**Figure 7.1:** Three isotope plot ($^{206}\text{Pb}/^{207}\text{Pb}$ vs $^{208}\text{Pb}/^{207}\text{Pb}$) of soil (closed circles) and bioaccessible (open circles) Pb in Glasgow, with mean isotope ratios for Broken Hill/Australian Ore, UK petrol (pre-1998), ore from Wanlockhead (in the Leadhills southwest Scotland), and UK coal. From Farmer et al., 2011.
\(^{206}\text{Pb}/^{207}\text{Pb}\) ratio in paint is 1.117, with a large range from 1.083 to 1.183 (Sugden et al., 1993). Consequently, it can be difficult to accurately quantify the proportion of Pb from multiple sources in the soil.

Recycling and waste management can also pose a problem, as approximately half the Pb used today is from recycled sources (Robinson, 2009). When modern Pb-containing materials e.g. car batteries and computer monitors, are deposited in landfills, Pb may enter the soil via leaching in acidic conditions, or incineration which will generate atmospheric Pb emissions (Robinson, 2009; Bond et al., 2013). These modern Pb-materials contain a combination of undistinguishable Pb isotopes, as Pb used in these materials may be recycled or originate from multiple ores. Overall, when this Pb enters the soil, source identification is very difficult.

Despite these limitations, several studies have examined the relationship between the isotopic composition of Pb and its sources in Scotland (Table 7.1). An early, notable study by Sudgen et al. (1993) examined isotopic Pb ratios in contemporary environmental materials in Edinburgh, including leaded petrol, atmospheric particles, house and street dust, paint, drinking water, and coal, but unfortunately not soil. They found 71\% of Pb in street dust originated from petrol emissions, with the remaining 29\% likely originating from old paint.

Farmer et al. (2011) calculated Pb isotope ratios in G-BASE samples from Greater Glasgow and found similar isotopic ratios to MacKinnon et al. (2011), despite much higher total Pb concentrations (Table 7.1). Farmer et al. (2011) aimed to investigate Pb bioaccessibility; therefore, did not explore the relationship between Pb isotopes and road proximity. However, the percentage of Pb from petrol can be calculated from their data and is similar to results from other studies in the Glasgow area (Table 7.1).

Other studies have examined the change in Pb isotopes with depth down the soil profile (Bacon and Hewitt, 2005; Farmer et al., 2006; Farmer et al., 2011). These studies observed lower \(^{206}\text{Pb}/^{207}\text{Pb}\) ratios in topsoil compared with higher \(^{206}\text{Pb}/^{207}\text{Pb}\) ratios deeper in the soil. This pattern likely occurs because more recent Pb originating from petrol will have been deposited closer to the surface, whereas further back in time; therefore, deeper in the soil profile, more Pb will have originated from coal or industrial sources which have higher \(^{206}\text{Pb}/^{207}\text{Pb}\) ratios. Therefore, the \(^{206}\text{Pb}/^{207}\text{Pb}\) ratio may be influenced by the sample depth, as noted in Table 7.1.
<table>
<thead>
<tr>
<th>Study</th>
<th>Mean (Min. – max.)</th>
<th>Samples</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Soil Pb mg kg⁻¹</td>
<td>206Pb/207Pb Petrol Pb % Petrol conc. mg kg⁻¹</td>
<td>Number</td>
</tr>
<tr>
<td>Bacon and Hewitt, 2005</td>
<td>168 (99 - 202)</td>
<td>1.149 (1.141 - 1.167) 31* (14 - 38) 52* (13 - 78)</td>
<td>5</td>
</tr>
<tr>
<td>Farmer et al., 2011</td>
<td>659 (126 - 2,160)</td>
<td>1.143 (1.057 - 1.175) 33* (6 - 83) 187* (0 - 709)</td>
<td>27</td>
</tr>
<tr>
<td>Farmer et al., 2016</td>
<td>77 (4 - 580)</td>
<td>1.156 (1.105 - 1.202) n/a n/a</td>
<td>103</td>
</tr>
<tr>
<td>Macintyre, 2017</td>
<td>312 (134 - 501)</td>
<td>1.155 (1.136 - 1.168) 25 (12 - 42) 84 (24 - 191)</td>
<td>18</td>
</tr>
<tr>
<td>MacKinnon et al., 2011</td>
<td>190 (49 - 703)</td>
<td>1.134 (1.091 - 1.165) 45* (15 - 86) 103* (14 - 569)</td>
<td>31</td>
</tr>
<tr>
<td></td>
<td>141 (59 - 282)</td>
<td>1.137 (1.097 - 1.167) 41* (13 - 80) 64* (9 - 215)</td>
<td></td>
</tr>
</tbody>
</table>
A study by MacKinnon et al. (2011) used isotope analysis on road transect soil samples from 15 km west of Glasgow, to apportion soil Pb to either a petrol or other (mainly industrial) sources. Overall, they observed decreasing soil Pb originating from petrol with increasing distance from the road (Figure 7.2), but no clear trend in industrial emissions, as these were dependent on historical inputs. They also compared transects away from roads with differing traffic densities and found that up to 10 m away from the M8 motorway soil Pb was elevated due to petrol sources, whereas soil Pb was only higher than baseline within 3 m of a minor road (Figure 7.2A and B). Lastly, the trend of decreasing soil Pb from petrol sources away from roads may become obscured over time, with samples from 2010 showing a more irregular decline compared with the smooth exponential decay exhibited by the 2001 samples (Figure 7.2C and D; MacKinnon et al., 2011).

**Figure 7.2:** Mean total Pb and Pb originating from petrol and non-petrol sources in soil samples collected away from (A) a motorway and (B) a minor road. (C) and (D) compare petrol Pb concentrations away from a motorway and minor road in 2001 and 2010. From MacKinnon et al., 2011
Similarly, Macintyre (2017) determined total Pb concentrations and the isotopic ratio of 42 samples from Richmond Park in central Glasgow, including 18 samples collected from two roadside transects. Macintyre used the $^{206}$Pb/$^{207}$Pb ratio with total Pb concentrations to determine the percentage and amount of Pb originating from petrol. In each roadside transect there is a clear exponential decrease in Pb from petrol away from the road (Figure 7.3), a pattern which is not as strong in total Pb concentrations. Furthermore, petrol emissions only enhanced soil Pb up to a maximum of 10 m from the road edge. This is comparable with MacKinnon et al.’s (2011) findings, despite differences in location, sampling dates, and road type.

![Figure 7.3: Total Pb, Pb originating from petrol, and Pb from non-petrol sources in soil samples collected with increasing distance from the same road in Richmond Park i.e. transects (A) and (B). From Macintyre, 2017.](image)

Overall, the research in Table 7.1 demonstrates that Pb isotope ratios can be used to apportion the source(s) of Pb in soil, with an average of 25% of Pb originating from petrol in most soils. However, isotopic ratios may be influenced by sampling depth, date, and location, which should be taken into consideration when comparing results.

7.2. **Chapter aims**

1. To determine the percentage and total concentration of soil Pb from petrol and coal sources in Paisley and Bishopbriggs.
2. To model and map soil Pb from petrol and coal sources.
3. To determine what is influencing the spatial distribution of soil Pb from petrol and coal sources.
7.3. **Methodology**

Lead isotopes ratios were determined by analysing the digested aqueous solutions (see Section 3.2.3.1.) with inductively coupled plasma mass spectroscopy (ICP-MS). All samples were analysed in duplicate. Before analysis, samples were diluted to Pb concentrations of <20 μgL⁻¹ using 2% HNO₃ (Aristar™). Lead isotope ratios (²⁰⁶Pb/²⁰⁷Pb, ²⁰⁸Pb/²⁰⁷Pb, and ²⁰⁸Pb/²⁰⁶Pb) were determined using pulse count mode on a single quadropole ICP-Q-MS (Agilent 7500ce), with octopole re-action system and nickel cones (Agilent Technologies, Stockport, UK). The argon gas flows were 0.81 and 0.22 L min⁻¹ for carrier and make-up flows, with a plasma gas flow of 15 L min⁻¹. The peristaltic pumping rate into the Miramist nebulizer was 0.06 rps or 1.2 ml min⁻¹. Simultaneous isotope analysis was employed, with one peak per Pb isotope, a two second integration time, and five replicate measurements per sample. A Pb isotopic reference material from the National Institute of Standards and Technology (SRM 981), was used for mass bias correction.

The ICP-MS instrument operates firstly by combining the aqueous sample and argon gas in a nebulizer, which converts the sample into aerosol particles (Thomas, 2003; PerkinElmer, 2011). Larger aerosol particles are then removed in the spray chamber and the smaller particles enter the high temperature (>6,000 K) inductively coupled plasma torch (Thomas, 2003). The plasma dries the aerosol sample, which is subsequently dissociated, atomised, and ionised to generate positive and neutral ions (Thomas, 2003). Next, these ions are separated by off-axis ion lenses in the interface, so that only positive ions remain. The positive ions are filtered through an octopole ion reaction system to remove background interferences, then a hyperbolic quadropole mass spectrometer uses voltages to create an electrostatic filter (Thomas, 2003). This filter controls which ions move through the quadropole based on their mass to charge ratio, which is specific to each Pb isotope (Thomas, 2003). Finally, the ions belonging to a specific isotope strike the detection system and generate an ionic signal pulse (Thomas, 2003). The intensity of this pulse is compared with the standard calibration curve to determine the concentration of each isotope (Thomas, 2003).

During analysis, a number of quality assurance checks were performed. Firstly, the Pb solutions were diluted to <20 ppb as this generates high enough counts to obtain accurate measurements, while ensuring the detector remains below the changeover point (i.e. <1,000,000 counts). This is important as samples containing
Pb close to this point would cause the detector to switch between pulse count mode and analogue mode, which would add error to the isotopic ratios. To ensure precision, five repeat measurements of each duplicate subsample were analysed; therefore, each ratio is an average of ten measurements, with standard deviations and confidence intervals calculated to check precision. Accuracy was assessed by carrying out mass bias corrections and calibrating the instrument with SRM981 standard solution (National Institute of Standards and Technology, diluted to 10 ppb and 20 ppb). Furthermore, an external in-house standard (SRM1640a) was analysed after calibration to double-check accuracy. Every six samples the SRM981 10 ppb and 20 ppb standards were reanalysed to check for drift. Four reagent blanks were also analysed per run to assess cross-contamination, and ratios were corrected for blank values during post-analysis. Variability between runs was assessed by reanalysing a random sample from the previous run. All chemicals used were of Aristar™ quality and only high-purity water was used.

Lead isotope ratios were determined for 48/138 samples in Bishopbriggs and 49/112 samples in Paisley, providing data on 97 samples in total. These samples were selected at random, and this number of samples were analysed due to time and financial constraints. Overall, this provides a rich dataset for Pb isotopic ratios in Paisley and Bishopbriggs, sufficient to address the project aims. To provide further information, the isotopic ratios were joined with total Pb results (from ICP-OES), sample coordinates, and external factors e.g. distance to road and historic industry, using a VLOOKUP table in Microsoft Excel. Subsequently, the percentages of Pb from petrol or coal sources were calculated using Equation 7.1 (adapted from Farmer et al. 1996):

\[
Pb_{\text{petrol}} = \frac{^{206}\text{Pb}^{/207}\text{Pb}_{\text{Coal}} - ^{206}\text{Pb}^{/207}\text{Pb}_{\text{Sample}}}{^{206}\text{Pb}^{/207}\text{Pb}_{\text{PETROL}} - ^{206}\text{Pb}^{/207}\text{Pb}_{\text{COAL}}} \times 100\% \quad (7.1)
\]

\[
Pb_{\text{coal}} = \frac{^{206}\text{Pb}^{/207}\text{Pb}_{\text{PETROL}} - ^{206}\text{Pb}^{/207}\text{Pb}_{\text{Sample}}}{^{206}\text{Pb}^{/207}\text{Pb}_{\text{PETROL}} - ^{206}\text{Pb}^{/207}\text{Pb}_{\text{COAL}}} \times 100\%
\]

Farmer et al. 1996): The reference $^{206}\text{Pb}^{/207}\text{Pb}$ ratios for petrol and coal were selected from the literature and were based on similar locations and the most recent measurements, as Pb isotopic ratios vary geographically and over time (Farmer et al. 1999). The percentage of Pb originating from petrol and coal sources was multiplied by total Pb to determine the concentration (mg kg$^{-1}$) of Pb from petrol and coal sources (Table 7.2).
Table 7.2: The five key variables used in this analysis, with shorthand notation, units, and calculation method detailed.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Known as</th>
<th>Unit</th>
<th>Calculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Pb</td>
<td>Total(_{Pb})</td>
<td>mg kg(^{-1})</td>
<td>Average Pb measurements from duplicate samples using ICP-OES</td>
</tr>
<tr>
<td>Percentage of Pb from petrol</td>
<td>Petro(_{Pb})</td>
<td>%</td>
<td>Equation 7.1(A)</td>
</tr>
<tr>
<td>Percentage of Pb from coal</td>
<td>Coal(_{Pb})</td>
<td>%</td>
<td>Equation 7.1(B)</td>
</tr>
<tr>
<td>Pb concentration from petrol</td>
<td>Petro(_{Pb, \text{conc}})</td>
<td>mg kg(^{-1})</td>
<td>Total(<em>{Pb}) * (Petro(</em>{Pb})/100)</td>
</tr>
<tr>
<td>Pb concentration from coal</td>
<td>Coal(_{Pb, \text{conc}})</td>
<td>mg kg(^{-1})</td>
<td>Total(<em>{Pb}) * (Coal(</em>{Pb})/100)</td>
</tr>
</tbody>
</table>

Each of the five variables in Table 7.2 were then explored in Paisley and Bishopbriggs. Firstly, summary statistics were calculated, which involved determining the mean, median, range (minimum and maximum values), standard deviation, and octile skew of each variable in Matlab. Subsequently, each variable was modelled using REML to fit a variogram (see Equation 5.8 for details). Lastly, each model was assessed by calculating its AIC and performing 10-fold cross validation (see Equations 5.10 and 6.1 for details).

To map each variable in Table 7.2, the variable in question was ordinary kriged (if no covariates were present) or E-BLUP was run, if covariates were included. Sections 5.3.3. and 6.3.2 (Equation 6.5) provide a more in-depth explanation of these processes. The kriged prediction grid was then exported into QGIS where each variable was mapped. Percentile breaks in the isotope ratio legend were calculated for each variable using all Paisley and Bishopbriggs measurements, as this allows direct comparison between maps.

Covariates were then explored to determine controls on the spatial patterns of the five key variables. Specifically, the relationship between each variable (Table 7.2) and each covariate i.e. road distance and historic industry distance, were investigated. This was achieved by generating a scattergraph of the modelled value of each variable versus road or industry distance, to give a general picture on their relationship. More in-depth analysis was carried out by remodelling each variable but with road or historic industry distance as a covariate (see Section 6.3.2.). The outputs from the covariate models were then compared with the simple i.e. no covariate, models. Furthermore, Wald values (Equation 6.4) were calculated to determine the strength of the relationship between each variable and each covariate.
7.4. Results and discussion

7.4.1. Quality control checks of isotopic results

Mean precision errors (±1 SD) from quintuplicate replicate measurements of duplicate sample ratios (n=97) were ±0.0024 for $^{206}\text{Pb}/^{207}\text{Pb}$, ±0.0045 for $^{208}\text{Pb}/^{207}\text{Pb}$, and ±0.0042 for $^{208}\text{Pb}/^{206}\text{Pb}$, which are low enough to be able to differentiate between Pb sources. Furthermore, the largest errors are only ±0.0041 for $^{206}\text{Pb}/^{207}\text{Pb}$, ±0.0077 for $^{208}\text{Pb}/^{207}\text{Pb}$, and ±0.013 for $^{208}\text{Pb}/^{206}\text{Pb}$. The relative standard deviations (RSD, i.e. SD/Mean *100) were also low at ±0.2% (±0.21%, ±0.19%, and ±0.2% for $^{206}\text{Pb}/^{207}\text{Pb}$, $^{208}\text{Pb}/^{207}\text{Pb}$, and $^{208}\text{Pb}/^{206}\text{Pb}$ respectively).

Additionally, the variation between duplicate samples was compared with the within sample difference (i.e. quintuplicate measurements). The majority of the samples i.e. >75%, showed no difference when comparing the within and between sample standard deviations.

To ensure analytical accuracy an external in-house standard (SRM 1640a) was analysed every run, and its mean isotope ratios (±1 SD) were 1.165±0.005, 2.447±0.006, and 2.1±0.006 (n = 4) for $^{206}\text{Pb}/^{207}\text{Pb}$, $^{208}\text{Pb}/^{207}\text{Pb}$, and $^{208}\text{Pb}/^{206}\text{Pb}$, respectively. These are in close agreement (<0.012 difference) with previously analysed values of 1.174±0.004, 2.459±0.005, and 2.095±0.007 (Farmer et al., 2016).

Additionally, drift checks were performed on the SRM 981 certified standard every six samples. The average instrumental drift was very close to 0% for all Pb isotope ratios (±0.0024%, ±0.028%, and ±0.043% for $^{206}\text{Pb}/^{207}\text{Pb}$, $^{208}\text{Pb}/^{207}\text{Pb}$, and $^{208}\text{Pb}/^{206}\text{Pb}$ respectively), with overall maximum variation at <1% (n = 49). Furthermore, the absolute difference between the certified and measured values does not follow any pattern and is generally very low with averages of ±0.0016, ±0.003, and 0.0025 for $^{206}\text{Pb}/^{207}\text{Pb}$, $^{208}\text{Pb}/^{207}\text{Pb}$, and $^{208}\text{Pb}/^{206}\text{Pb}$, respectively.

To assess variability between runs, one random sample from the previous run was reanalysed in the following run. The average difference between repeat samples (n = 4) was low overall, at ±0.007, ±0.006, and ± 0.01 for $^{206}\text{Pb}/^{207}\text{Pb}$, $^{208}\text{Pb}/^{207}\text{Pb}$, and $^{208}\text{Pb}/^{206}\text{Pb}$, respectively.
7.4.2. Percentage and total Pb from petrol and coal sources

Table 7.3 summarizes the isotopic ratios, Pb percentages, and total amounts of Pb from petrol and coal sources in Paisley and Bishopbriggs. When compared with other studies (Table 7.1) the $^{206}\text{Pb}/^{207}\text{Pb}$ ratios in Paisley are very similar; however, they are slightly higher in Bishopbriggs, and the range is narrower in both areas, particularly with fewer low values. This could occur because the samples in this study were collected more recently (Table 7.1); therefore, Pb from leaded-petrol emissions (banned in 2000), may have had more time to dissipate. Furthermore, this study did not focus on road transects like other studies (e.g. MacKinnon et al. 2011; Macintyre, 2017), thus will include more samples further from roads. These samples will likely have accumulated more Pb from coal, paint, or background Pb emissions and consequently will have higher $^{206}\text{Pb}/^{207}\text{Pb}$ ratios.

| Table 7.3: Descriptive statistics of isotope ratios and key variables (Table 7.2) for (A) Paisley and (B) Bishopbriggs. S.D. = standard deviation; O.S. = octile skew. |
|---|---|---|---|---|---|---|---|
|  | $^{206}\text{Pb}/^{207}\text{Pb}$ | $^{208}\text{Pb}/^{207}\text{Pb}$ | Total Pb conc. (mg kg$^{-1}$) | Coal Pb conc. (mg kg$^{-1}$) | Petrol Pb conc. (mg kg$^{-1}$) | Coal Pb % | Petrol Pb % |
| A - PAISLEY | Count | 49 | 49 | 49 | 49 | 49 | 49 | 49 |
|  | Mean | 1.147 | 2.43 | 516 | 316.53 | 199.48 | 65.92 | 34.08 |
|  | Median | 1.148 | 2.43 | 289.33 | 198.73 | 93.79 | 66.67 | 33.33 |
|  | Min. | 1.124 | 2.405 | 51.65 | 39.45 | 9.13 | 42.42 | 16.67 |
|  | Max. | 1.165 | 2.45 | 1765.7 | 1177.1 | 689.46 | 83.33 | 57.58 |
|  | S.D. | 0.01 | 0.011 | 488.35 | 300.29 | 198.46 | 10.15 | 10.15 |
|  | O.S. | -0.11 | -0.065 | 0.59 | 0.512 | 0.67 | -0.11 | 0.11 |
| B - BISHOPBRIGGS | Count | 48 | 48 | 48 | 48 | 48 | 48 |
|  | Mean | 1.16 | 2.444 | 430.09 | 340.53 | 89.57 | 78.48 | 21.52 |
|  | Median | 1.161 | 2.445 | 166.08 | 128.72 | 36.48 | 80.05 | 19.95 |
|  | Min. | 1.133 | 2.418 | 75.19 | 58.481 | 9.45 | 51.01 | 10.10 |
|  | Max. | 1.171 | 2.458 | 2422.2 | 2006.3 | 415.94 | 89.9 | 48.99 |
|  | S.D. | 0.0079 | 0.009 | 547.84 | 446.68 | 106.98 | 8.01 | 8.01 |
|  | O.S. | -0.24 | -0.17 | 0.85 | 0.87 | 0.80 | -0.24 | 0.24 |
The percentage and amount of soil Pb originating from petrol in Paisley and Bishopbriggs is similar to other research (Tables 7.1 and 7.3). Specifically, the Bishopbriggs mean petrol concentration is low at 90 mg kg\(^{-1}\), but it is comparable with Mackinnon et al.’s (2011) mean value of 64 mg kg\(^{-1}\), and Macintyre’s (2017) findings of mean petrol soil Pb at 84 mg kg\(^{-1}\). Paisley’s higher petrol Pb mean of 199 mg kg\(^{-1}\) is akin to Farmer et al.’s (2011) measurement of soil Pb from petrol sources at 187 mg kg\(^{-1}\). As previously discussed, it is difficult to compare Pb isotope ratios from different studies as the ratios will be influenced by several factors, e.g. dilution and sampling strategies. However, it is reassuring that Pb isotope ratios measured in this study are comparable to other research in similar locations.

The isotope plots (Figure 7.4) show that all samples lie close to a tie line (and within the error range) of petrol and coal Pb sources. Therefore, Pb in the soil samples can be attributed to a mix of petrol and coal origins. The plot also illustrates that soil samples from Paisley have a slightly lower \(^{206}\text{Pb}/^{207}\text{Pb} and \(^{208}\text{Pb}/^{207}\text{Pb}\) ratios (Figure 7.4), suggesting that more Pb originates from petrol in Paisley compared with Bishopbriggs. Furthermore, in Bishopbriggs many samples have high (>1.16 to 2.44) \(^{206}\text{Pb}/^{207}\text{Pb} and \(^{208}\text{Pb}/^{207}\text{Pb}\) ratios (Figure 7.4), implying a strong coal influence on Pb in this area. This is unsurprising, as coal mining was extensive in Bishopbriggs in the 19\(^{th}\) and early 20\(^{th}\) centuries (Scottish Mining, 2011).
Figure 7.4: Three Pb isotope plot ($^{206}Pb/^{207}Pb$ vs $^{208}Pb/^{207}Pb$) of soil Pb from Bishopbriggs (green) and Paisley (blue) samples, dot = mean ratio, error lines = +/- one standard deviation from repeat measurements. Insert is shown in the bottom plot. Reference isotope ratios (red) are also depicted for UK petrol until 1998, ore from Wanlockhead, and Scottish coal (Farmer et al. 2011), with the tie line and error range between petrol and coal shown as grey shading.
In both Paisley and Bishopbriggs most Pb originates from coal sources (mean = 66% in Paisley and 78% in Bishopbriggs, Table 7.3; Figure 7.4). Interestingly, coal Pb concentrations have similar mean values in Paisley and Bishopbriggs, 317 mg kg\(^{-1}\) and 341 mg kg\(^{-1}\) respectively (Table 7.3). This confirms the earlier theory that background Pb emissions, from the burning of coal, were relatively consistent across Glasgow (see Section 4.4.4.).

By contrast, mean petrol concentrations are much lower in Bishopbriggs compared with Paisley, 90 mg kg\(^{-1}\) and 199 mg kg\(^{-1}\) respectively (Table 7.3). This suggests that petrol emissions show more variation over both locations, than coal emissions. Lower petrol Pb emissions in Bishopbriggs could have occurred for several reasons.

Firstly, more samples collected further from roads would result in lower petrol Pb concentrations. However, this is not the case as the mean distance to road is 19 m in Bishopbriggs and 20 m in Paisley, with the majority of samples occurring <20 m from a road in both locations (62% in Bishopbriggs and 63% in Paisley).

Another explanation could be that the roads in Bishopbriggs were constructed more recently, thus have experienced a shorter history of leaded petrol emissions. Again, this is not applicable as historic maps show that roads in Paisley were constructed from pre-1896 to recently (2000s), as were roads in Bishopbriggs.

Other studies have observed that generally, roads with higher traffic densities pollute roadside soils to a greater degree and for a further distance (e.g. Welch and Dick, 1975; Wiseman et al., 2013; Azeez et al., 2014). Historic traffic density data is difficult to obtain, but currently Springfield Road in Bishopbriggs has around 4,000 vehicles per year and Barskiven Road in Paisley 7,000 vehicles per year (Road Traffic Stats.gov). The higher traffic density in Paisley could explain the higher amount of Pb from petrol, compared with Bishopbriggs.

Soil properties are also likely to have influenced the extent of Pb elevation in roadside soils. Specifically, an increased sorption capacity has been observed in some roadside soils with a higher pH from calcareous materials, higher organic carbon levels, more clay minerals, and more black carbon from tyre wear (Kluge and Wessolek, 2012). Therefore, further investigation into the soil properties around roads in Paisley and Bishopbriggs is required.
7.4.3. Modelling and mapping petrol and coal soil Pb

Table 7.4 outlines the variogram parameters and validation statistics for each key variable model in Paisley and Bishopbriggs. Overall, the variograms have a low nugget and higher partial sill (Table 7.4A), indicating that spatial autocorrelation is present in the samples and that the sampling strategy was sufficient to capture it. Variograms of coal Pb percentages have exceptionally high semivariance values due to the high lambda conversion values (i.e. 3.28 and 1.59). The effective range is consistent with earlier research (i.e. 72 m and 116 m for Paisley and Bishopbriggs, Table 5.5) and suggests that soil Pb varies up to ~80 m, with slightly larger ranges in petrol soil Pb (66 - 98 m) compared with coal soil Pb (55 - 86 m), and in Bishopbriggs compared with Paisley (68 - 98 m vs 55 - 87 m; Table 7.4A).

The cross-validation results reveal that the models are relatively effective at estimating Pb concentrations and percentages, as the mean SSE does reach 1.3 but is generally close to 1, and the median SSE is as high as 0.74 but overall is close to 0.45 (Table 7.4B). The bias/mean error and RMSE is worryingly high (>50 or 200) in some models i.e. high RMSE in total and coal Pb concentrations in Bishopbriggs, and in total, petrol, and coal Pb concentrations in Paisley (Table 7.4B). Overall, it is difficult to compare models for different variables in separate locations, but the percentage Pb models appear to be more accurate, possibly due to less skew in these datasets (Table 7.3).

| Table 7.4: (A) Variogram properties and (B) validation statistics from for the five key variables models in Paisley and Bishopbriggs, with no covariates. RMSE in original units. |
|---|---|---|---|---|---|---|---|---|
| **Pb conc.** | **Pb Percent** | **Pb conc.** | **Pb Percent** |
| Total | Petrol | Coal | Total | Petrol | Coal |
| **BISHOPBRIGGS** | | | | |
| Samples | 48 | 48 | 48 | 48 | 48 | 49 | 49 | 49 | 49 | 49 |
| Nugget | 0 | 0.01 | 0 | 0 | 0 | 0 | 0 | 0 | 1.10 | 4703 |
| Partial sill | 0.004 | 0.11 | 0.001 | 0.01 | 2.05x10^{-10} | 1.65 | 3.85 | 0.75 | 2.24 | 8968 |
| Eff. range (m) | 81 | 98 | 85 | 68 | 68 | 62 | 66 | 55 | 87 | 86 |
| Smoothness | 1.5 | 1.50 | 1.50 | 1.50 | 0.43 | 1.50 | 1.50 | 1.50 | 1.50 |
| Lambda | -0.49 | -0.27 | -0.54 | -0.38 | 3.28 | 0.03 | 0.09 | -0.03 | 0.51 | 1.59 |
| **PAISLEY** | | | | |
| Samples | | | | |
| Nugget | | | | |
| Partial sill | | | | |
| Eff. range (m) | | | | |
| Smoothness | | | | |
| Lambda | | | | |

| B - Validation statistics (non-transformed units) |
|---|---|---|---|---|---|---|---|
| Likelihood | 332 | 251 | 308 | 1589 | 161 | 349 | 303 | 324 | 180 | 180 |
| ME/Bias | 63 | 11 | 47 | -0.2 | 0.2 | 0.1 | -10 | 14 | 0 | 0.5 |
| RMSE | 342 | 63 | 248 | 6 | 6 | 330 | 139 | 249 | 8 | 8 |
| Mean SSE | 1.24 | 1.02 | 1.11 | 1.11 | 1.28 | 0.86 | 0.94 | 1.30 | 1.26 | 1.24 |
| Median SSE | 0.6 | 0.74 | 0.53 | 0.39 | 0.4 | 0.62 | 0.56 | 0.74 | 0.71 | 0.62 |
| AIC | 651 | 513 | 628 | 329 | 334 | 701 | 6156 | 661 | 371 | 372 |
In Paisley, the ordinary kriged maps of total and percentage of Pb from petrol or coal sources (Figure 7.5) show some unexpected patterns.

Firstly, although the coal Pb percentage hotspots correlate with the locations of historical industry, such as the old steel works in the southwest and the refuse heap in the northeast (Figure 7.5E), these hotspots are relatively localised and do not translate onto the soil Pb concentration from coal sources map (Figure 7.5C). In fact, the hotspots on the coal percentage map (Figure 7.5E) are low spots on the coal Pb concentration map (Figure 7.5C). This is because the coal Pb percentage hotspot which centres on sample 11, has a very high percentage of Pb from coal sources (82%, Figure 7.5E), but low total Pb levels (109 mg kg\(^{-1}\), Figure 7.5A), which results in only 90 mg kg\(^{-1}\) of total Pb from coal sources (Figure 7.5C). Furthermore, the hotspot of soil Pb concentration from coal sources is slightly east of the percentage coal Pb hotspot (Figure 7.5C/E). In the soil Pb concentration from coal sources hotspot, samples such as 38 have low relative percentages of Pb from coal (~50% which is <5\(^{th}\) percentile, Figure 7.5E), but extremely high overall soil Pb levels (almost 1,000 mg kg\(^{-1}\), which is >95\(^{th}\) percentile, Figure 7.5A). This results in approximately 500 mg kg\(^{-1}\) of soil Pb from coal sources, which is between the 75\(^{th}\) and 95\(^{th}\) percentile (Figure 7.5C). This illustrates that the total soil Pb concentration often dominates over the actual percentage of soil Pb from coal sources.

Soil Pb from petrol in Paisley does correlate with the road distribution in some places, such as around sample 76 (Figure 7.5B/D). However, there are roads surrounded by low total soil Pb from petrol, such as sample 49 (Figure 7.5B/D). The spatial distribution of the percentage of Pb from petrol (Figure 7.5D) generally correlates with total soil Pb from petrol sources map (Figure 7.5B). Overall, the spatial distribution of total soil Pb from petrol or coal sources is largely driven by total soil Pb. However, the percentage of soil Pb from petrol or coal sources paints a different picture.
Figure 7.5: Kriged soil Pb percentage and concentration from petrol and coal sources in Paisley, with total Pb for comparison (based only on the 49 samples with isotope data). Generated from REML modelled values with no covariates. Categorised into six percentiles based on all samples from Paisley and Bishopbriggs values to allow comparison between locations. Specific samples, roads, and historic industry locations also depicted for reference.
In Bishopbriggs the soil Pb spatial distribution is largely controlled by coal emissions (Figure 7.6). Unsurprisingly, the distribution of soil Pb concentrations from coal sources (Figure 7.6C), closely corresponds to overall soil Pb (Figure 7.6A). However, the map of the percentage of Pb from coal reveals some interesting patterns. Firstly, there are high levels and percentages of coal Pb around Etive Crescent Park in the NE (i.e. sample 73 = 80% Pb from coal, ~1,725 mg kg\(^{-1}\) total coal Pb), which supports the hypothesis that the refuse heap was filled with industrial waste, likely from the nearby coal mine (Figure 7.6E). There are also high percentages of Pb from coal sources around the houses to the south i.e. sample 112 where coal Pb is 86% (Figure 7.6E). These are newer (post-1980) houses, thus the coal Pb in the surrounding soil likely originated from the nearby coal mine and not domestic sources. However, total levels of Pb are low in these samples (i.e. <400 mg kg\(^{-1}\), Figure 7.6A) which generates low total coal soil Pb concentrations (Figure 7.6C). Another small hotspot of the percentage of coal Pb centres around sample 101 (89% coal Pb, Figure 7.6E), which may relate to the nearby railway or old quarry, although this will depend on the type of material mined from the quarry, and type of coal used in the railway.

Petrol Pb levels are generally low in Bishopbriggs, as previously discussed. However, there is a hotspot of petrol Pb percentage around sample 100, where almost 50% of Pb originates from petrol (Figure 7.6D). This sample is located on a square surrounded by early 20\(^{th}\) century buildings, thus it follows that the roads were constructed at a similar time, so should have experience a long history of leaded petrol emissions. However, total soil Pb from petrol levels are relatively low (~200 mg kg\(^{-1}\)) and very localised.

Lastly, it is interesting that despite the percentage of Pb from petrol being low around the refuse heap hotspot (15-20% e.g. sample 73, Figure 7.6D), total soil Pb from petrol is relatively high (e.g. 400 mg kg\(^{-1}\), Figure 7.6B) due to high overall Pb levels (e.g. 2,000 mg kg\(^{-1}\), Figure 7.6A). This may indicate that Pb in the refuse heap could originate from multiple sources, and not just the nearby coal mine. This agrees with the earlier discussion that Pb isotopes have difficulty in distinguishing Pb provenance when multiple sources are mixed e.g. in a refuse heap (Robinson, 2009).
**Figure 7.6:** Kriged soil Pb percentage and concentration from petrol and coal sources in Bishopbriggs, with total Pb for comparison (based only on the 48 samples with isotope data). Generated from REML modelled values with no covariates. Categorised into six percentiles based on all samples from Paisley and Bishopbriggs values to allow comparison between locations.
7.4.4. The influence of road distance on petrol and coal soil Pb

In order to determine what is influencing the spatial pattern in soil Pb from petrol and coal sources, road distance was included as a covariate in the REML models. The results are shown in Table 7.5, and Figures 7.7 and 7.8.

**Table 7.5:** (A) Variogram properties, (B) validation statistics, and (C) correlation with soil Pb, from modelled values for the key variables in Paisley and Bishopbriggs with road distance as a covariate. Validation statistics are shaded green if they are an improvement on the no covariate model (Table 7.4). Significant with a **>=99% confidence *>=95% confidence.**

<table>
<thead>
<tr>
<th></th>
<th>BISHOPBRIGGS</th>
<th>PAISLEY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A - Variogram properties</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Samples</td>
<td>48</td>
<td>49</td>
</tr>
<tr>
<td>Nugget</td>
<td>0</td>
<td>0.00</td>
</tr>
<tr>
<td>Partial sill</td>
<td>0.01</td>
<td>0.14</td>
</tr>
<tr>
<td>Eff. range (m)</td>
<td>52.15</td>
<td>56.14</td>
</tr>
<tr>
<td>Smoothness</td>
<td>1.5</td>
<td>1.50</td>
</tr>
<tr>
<td>Lambda</td>
<td>-0.4</td>
<td>-0.21</td>
</tr>
<tr>
<td><strong>B - Validation statistics</strong> (non-transformed units)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Likelihood</td>
<td>323.3</td>
<td>245.53</td>
</tr>
<tr>
<td>ME/Bias</td>
<td>32.97</td>
<td>9.02</td>
</tr>
<tr>
<td>RMSE</td>
<td>260.9</td>
<td>57.57</td>
</tr>
<tr>
<td>Mean SSE</td>
<td>1.15</td>
<td>1.23</td>
</tr>
<tr>
<td>Median SSE</td>
<td>0.44</td>
<td>0.58</td>
</tr>
<tr>
<td>AIC</td>
<td>632.59</td>
<td>505.05</td>
</tr>
<tr>
<td><strong>C - Correlation statistics</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Wald</td>
<td>24.32</td>
<td>15.79</td>
</tr>
<tr>
<td>$p$</td>
<td>&lt;0.01**</td>
<td>&lt;0.01**</td>
</tr>
</tbody>
</table>

206
Figure 7.7: Scatterplots comparing road distance with (A) total and petrol Pb (unmodelled), (B) petrol Pb concentration, (C) petrol Pb percentage, (D) coal Pb concentration, and (E) coal Pb percentage, in Paisley. Plots B to E are for Box-Cox transformed (Pb) and modelled data, with road distance included as a covariate.
**Figure 7.8:** Scatterplots comparing road distance with (A) total and petrol Pb (unmodelled), (B) petrol Pb concentration, (C) petrol Pb percentage, (D) coal Pb concentration, and (E) coal Pb percentage, in Bishopbriggs. Plots B to E are for Box-Cox transformed (Pb) and modelled data, with road distance included as a covariate.
In Paisley, there is a slightly lower partial sill and range in the models including road distance as a covariate (i.e. 0.57 - 11156 for the partial sill, and 48 - 77 m for the effective range; Table 7.5A), compared with the no covariate models (0.75 - 8968 for the partial sill, 55 - 87 m for the effective range; Table 7.4A). This suggests that road distance is accounting for some of the variation in soil Pb. However, there is less difference between the models in Bishopbriggs, apart from slightly shorter ranges for the Pb concentration road distance models (52 - 78 m when road distance is included Table 7.5A, vs 68 - 81 m with no covariate Table 7.4A). Overall, it is difficult to directly compare models because slightly different lambda values were used.

In Paisley, the majority of the cross-validation statistics (e.g. mean SSE, median SSE, AIC) are improved by including road distance as a covariate (i.e. green highlighted cells in Table 7.5B) Furthermore, the coal models have a greater improvement when road distance is included, compared with the petrol models. In Bishopbriggs, the Pb percentage models see no marked improvement in mean SSE in when including road distance as a covariate (Table 7.5B). However, the petrol and coal Pb concentration models do see a general improvement when road distance is included.

In Paisley there is a significant relationship between total, petrol, and coal soil Pb concentrations, with relatively low, but significant at a 95% confidence interval, Wald values (Wald = 4.67 - 4.74, \( p = 0.03 \); Table 7.5C). However, this relationship is weaker than in Bishopbriggs, where the Wald values are much higher (Wald = 15.8 - 24.3, \( p < 0.01 \); Table 7.5C). Despite the weaker relationship, the relationship between soil Pb from petrol and road distance is negative in Paisley (Figures 7.7B and C), which fits with previous research (Section 6.4.3.4; Table 6.3).

In Bishopbriggs, total, petrol, and coal Pb concentrations all strongly correlate with road distance, as proven by the significant and high Wald values (Wald = 15.8 - 24.3, \( p < 0.01 \); Table 7.5C). However, there is a positive relationship between road distance and petrol Pb concentration (Figure 7.8B). This is unexpected and does not fit with previous studies. As discussed previously (Section 6.4.4.), this positive relationship is largely driven by high petrol and coal Pb concentrations near the old refuse heap in Etive Crescent park, which is far away from any roads. However, the percentage of Pb from petrol does show a negative relationship with road distance (Figure 7.8C), despite not being significant (Table 7.5C). This suggests road
distance and soil Pb should have an overall negative relationship in Bishopbriggs, but this is skewed by the high Pb concentrations in the refuse heap.

Lead from coal sources (both the concentration and percentage) exhibit a generally positive trend in both Paisley (Figures 7.7D and E) and Bishopbriggs (Figures 7.8D and E). This is to be expected as further from roads coal Pb, likely from background industrial atmospheric emissions, is the dominant source of Pb.

When all 138 or 112 samples were modelled i.e. not just samples with isotope data, the relationship between total soil Pb concentration and road distance was weaker in Bishopbriggs (all samples Wald = 15.2, isotope samples Wald = 24.32) but similar in Paisley (all samples Wald = 7.7, isotope samples Wald = 4.74; see Table 6.11D and Section 6.4.2.). You would expect these results to be similar, as isotope measurements were performed on a random subsample of data. This may indicate that more isotopic measurements are required in Bishopbriggs.
7.4.5. The influence of historic industry distance on petrol and coal soil Pb

The same methodology was used but historic industry distance was included as a covariate in the REML models. The results are shown in Table 7.6, and Figures 7.9 and 7.10.

<table>
<thead>
<tr>
<th></th>
<th>BISHOPBRIGGS</th>
<th>PAISLEY</th>
<th>BISHOPBRIGGS</th>
<th>PAISLEY</th>
<th>BISHOPBRIGGS</th>
<th>PAISLEY</th>
<th>BISHOPBRIGGS</th>
<th>PAISLEY</th>
<th>BISHOPBRIGGS</th>
<th>PAISLEY</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean Pb (mg kg⁻¹)</strong></td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td><strong>Petrol Pb conc. (mg kg⁻¹)</strong></td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>6.16</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.81</td>
</tr>
<tr>
<td><strong>Coal Pb conc. (mg kg⁻¹)</strong></td>
<td>0.11</td>
<td>0.00</td>
<td>0.01</td>
<td>3.2x10⁶</td>
<td>1.79</td>
<td>3.74</td>
<td>0.82</td>
<td>1.67</td>
<td>27208.71</td>
<td>13561.76</td>
</tr>
<tr>
<td><strong>Petrol Pb %</strong></td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>48</td>
<td>49</td>
<td>49</td>
<td>49</td>
<td>49</td>
</tr>
<tr>
<td><strong>Coal Pb %</strong></td>
<td>0.0035</td>
<td>0.11</td>
<td>0.00</td>
<td>0.01</td>
<td>3.2x10⁶</td>
<td>1.79</td>
<td>3.74</td>
<td>0.82</td>
<td>1.67</td>
<td>27208.71</td>
</tr>
<tr>
<td><strong>Eff. range (m)</strong></td>
<td>80.81</td>
<td>95.72</td>
<td>84.32</td>
<td>66.36</td>
<td>68.16</td>
<td>61.51</td>
<td>65.44</td>
<td>54.06</td>
<td>87.23</td>
<td>86.95</td>
</tr>
<tr>
<td><strong>Smoothness</strong></td>
<td>1.5</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>0.42</td>
<td>1.5</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
<td>1.50</td>
</tr>
<tr>
<td><strong>Lambda</strong></td>
<td>-0.49</td>
<td>-0.27</td>
<td>-0.54</td>
<td>-0.37</td>
<td>3.33</td>
<td>0.04</td>
<td>0.09</td>
<td>-0.02</td>
<td>0.47</td>
<td>1.72</td>
</tr>
</tbody>
</table>

**A - Variogram properties**

**B - Validation statistics** (non-transformed units)

<table>
<thead>
<tr>
<th></th>
<th>Likelihood</th>
<th>ME/Bias</th>
<th>RMSE</th>
<th>Mean SSE</th>
<th>Median SSE</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>331.74</td>
<td>30.65</td>
<td>36.07</td>
<td>1.35</td>
<td>0.62</td>
<td>649.48</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>250.65</td>
<td>6.42</td>
<td>36.07</td>
<td>1.29</td>
<td>0.85</td>
<td>515.29</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>307.98</td>
<td>67.41</td>
<td>251.89</td>
<td>1.12</td>
<td>0.72</td>
<td>629.95</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>158.57</td>
<td>5.93</td>
<td>5.97</td>
<td>1.17</td>
<td>0.44</td>
<td>331.15</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>160.68</td>
<td>0.11</td>
<td>0.24</td>
<td>1.28</td>
<td>0.48</td>
<td>335.35</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>160.68</td>
<td>0.24</td>
<td>0.46</td>
<td>0.96</td>
<td>0.64</td>
<td>372.63</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>348.33</td>
<td>-20.35</td>
<td>-19.55</td>
<td>0.99</td>
<td>0.61</td>
<td>682.66</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>301.74</td>
<td>-10.55</td>
<td>-19.55</td>
<td>0.90</td>
<td>0.58</td>
<td>617.47</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>324.16</td>
<td>-19.55</td>
<td>200.71</td>
<td>0.96</td>
<td>0.64</td>
<td>662.32</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>179.31</td>
<td>7.79</td>
<td>7.79</td>
<td>1.24</td>
<td>0.50</td>
<td>372.63</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>179.66</td>
<td>7.57</td>
<td>7.57</td>
<td>1.21</td>
<td>0.71</td>
<td>373.32</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>179.66</td>
<td>0.49</td>
<td>0.49</td>
<td>1.21</td>
<td>0.71</td>
<td>373.32</td>
</tr>
</tbody>
</table>

**C - Correlation statistics**

<table>
<thead>
<tr>
<th></th>
<th>Wald</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>0.05</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>0.12</td>
<td>0.73</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>0.03</td>
<td>0.86</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>0.04</td>
<td>0.84</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>0.05</td>
<td>0.55</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>0.45</td>
<td>0.50</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>0.34</td>
<td>0.56</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>0.74</td>
<td>0.39</td>
</tr>
<tr>
<td><strong>BISHOPBRIGGS</strong></td>
<td>0.66</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>PAISLEY</strong></td>
<td>0.66</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Figure 7.9: Scatterplots comparing historic industry distance with (A) total and historic industry Pb (unmodelled), (B) petrol Pb concentration, (C) petrol Pb percentage, (D) coal Pb concentration, and (E) coal Pb percentage, in Paisley. Plots B to E are for box-cox transformed (Pb) and modelled data, with historic industry distance included as a covariate.
Figure 7.10: Scatterplots comparing historic industry distance with (A) total and historic industry Pb (unmodelled), (B) petrol Pb concentration, (C) petrol Pb percentage, (D) coal Pb concentration, and (E) coal Pb percentage, in Bishopbriggs. Plots B to E are for Box-Cox transformed (Pb) and modelled data, with historic industry distance included as a covariate.
In Bishopbriggs the variogram properties are not changed when historic industry distance is included as a covariate (Table 7.6A), compared with the no covariate models (Table 7.4A). This indicates that historic industry distance does not account for any of the variation in soil Pb.

In Paisley there is a slight change to the variogram properties when historic industry distance is included as a covariate (Table 7.6A). Specifically, the effective range is slightly reduced in the coal models for most variables (54 - 87 m for coal models, 55 - 87 m for no covariate models), and the partial sill is slightly lower (0.82 - 27209 for coal models, 0.75 - 8968 for no covariate models). This indicates that a small amount of the soil Pb variation is accounted for by historic industry distance. As discussed previously, different lambda values were used which makes it difficult to compare models.

The validation statistics are generally not improved when historic industry distance is included in the models in Bishopbriggs, and they are only minorly improved for the Paisley models (see green highlighted cells in Table 7.6B). In particular, the mean SSE is improved in all Paisley historic industry models, but other validation statistics are worse when historic industry is included as a covariate (Table 7.6B).

The correlation statistics indicate that there is no relationship between total, petrol, and coal Pb concentrations and percentages in Bishopbriggs and Paisley. As proven by the very low and non-significant Wald values (Wald = 0.03 - 0.74, p = 0.39 - 0.86; Table 7.6C). The scatterplots (Figures 7.9 and 7.10) agree with these findings, as the spread is wide and there is no clear trend. However, there appears to be a slight negative trend i.e. decreasing soil Pb away from historical industry, in non-modelled values (Figures 7.9A and 7.10A). Although there are too many 0 m distances with a wide range of soil Pb values to generate a significant relationship.

The lack of relationship between soil Pb and historic industry distance was also observed for the whole dataset (Wald = 0 in Paisley, Wald = 3.2 in Bishopbriggs; Table 6.11D). This was attributed to only specific types of historic industry generating high levels of soil Pb, and these were absent from the study areas (apart from the refuse heap in Etive Crescent park). Furthermore, soil Pb pollution away from a historic industry location will not be omni-directional; it will be influenced by slope and wind direction. It was hoped that soil Pb originating from coal sources would have a stronger relationship with historic industry distance, however historic industrial pollution was likely widespread throughout both study areas, due to the
burning of coal for domestic heating. Therefore, soil Pb from coal may be more related to building age than a specific type or distance to historic industry.

7.5. Conclusions

This research has proven that Pb isotopes can be used to apportion soil Pb into Pb originating from petrol versus Pb originally from coal sources. Furthermore, the percentage and concentration of soil Pb from petrol or coal can be calculated and mapped, with both displaying different spatial patterns.

In Paisley and Bishopbriggs most soil Pb originates from coal, with mean percentages of soil Pb from industrial sources at 66% and 78% respectively, and mean industry soil Pb concentrations of 316 mg kg\(^{-1}\) and 340 mg kg\(^{-1}\) (Table 7.3). Coal soil Pb likely originated from coal mining activities and waste, but also background emissions from coal burning as an industrial and domestic heat source and dumping of ash waste to fertilise garden soil.

Soil in Paisley has a higher percentage and concentration of Pb from petrol, compared with Bishopbriggs (34% vs 22% and 199 mg kg\(^{-1}\) vs 78 mg kg\(^{-1}\), Table 7.3). This is possibly due to differences in traffic density, and roadside soil conditions influencing soil Pb mobility.

When mapped, petrol soil Pb hotspots loosely correlate with the locations of roads, and high values of coal Pb correspond with historic industry locations (especially the coal mine and refuse heap in Bishopbriggs, Figures 7.5 and 7.6). However, this pattern is not consistent, indicating other factors are influencing the distribution of soil Pb from petrol or coal i.e. general background industrial emissions, older houses burning coal fires and disposing of ash on garden soil, and the use of leaded paint. Furthermore, the isotopes only relate to the topsoil which may have been disturbed by more recent activity e.g. house building. Soil properties, and subsequently Pb mobility, are also important but have not been accounted for.

There is a significant relationship between total Pb, petrol Pb, and coal Pb concentrations, with road distance. These relationships are much stronger in Bishopbriggs compared with Paisley. Unexpectedly, road distance and petrol Pb have a positive correlation in Bishopbriggs, which is largely driven by a refuse heap generating soil Pb pollution which is far from roads. However, the road distance and petrol soil Pb trend negative in Paisley. This fits with previous research, as
historic leaded-petrol emissions will have elevated Pb levels in roadside soil. Lastly, coal soil Pb and road distance show a positive correlation in both areas, as soil further from roads is more likely to originate from coal sources, compared with petrol sources.

There is no relationship between historic industry distance and coal, petrol, and total Pb in either location, which agrees with earlier research (Chapter 6). The lack of relationship may exist because coal Pb pollution was so widespread as background emissions, that it was not limited to specific locations or historic industries. Furthermore, not all types of historic industry will have generated Pb pollution, so the examples of historic industry in Paisley and Bishopbriggs may be too specific. The wind direction and slope may also play a role in the spread of Pb pollution away from a historic industry location. Overall, the relationship between historic industry and soil Pb is multifaceted; therefore, more examples are required to generalise, or a specific example needs to be studies in greater detail. It is difficult to explore these complex relationships with only 97 samples, although this is a lot more than most other studies of a similar nature (Table 7.1).

The analysis in this chapter does not make the case for road distance or historic industry distance to be included as covariates in the neighbourhood machine learning models. In fact, it highlights the complex and unique relationship present between soil Pb (both total and from different sources) and the covariates in each neighbourhood examined in this study. This emphasises the importance of fully understanding the relationship between covariates and soil Pb, before including them in machine learning models.
References - Chapter 7


Road traffic stats.gov Traffic Statistics manual count points. Accessed online on 05/05/2022 from: https://roadtraffic.dft.gov.uk/manualcountpoints/967611 and https://roadtraffic.dft.gov.uk/manualcountpoints/805401 accessed on 05/05/2022


Scottish Mining (2011). Accessed online on 09/05/2022, from: http://www.scottishmining.co.uk/390.html


Chapter 8. Machine learning prediction of soil Pb at different scales.

Previous chapters have used geostatistics to identify that soil Pb concentrations are heterogenous across all study areas in Glasgow (Chapter 5), and to formally test which covariates have a significant and justifiable relationship with soil Pb concentrations (Chapter 6) and isotopic ratios (Chapter 7). However, this was limited to considering linear relationships between soil Pb and covariates. Machine learning methods can accommodate more complex relationships between a soil property and multiple covariates. Therefore, they have become increasingly popular at predicting the distribution of a soil property based on covariate values.

This chapter will explore how machine learning can be used to predict soil Pb concentrations, and the probability that these will be high, at the city and neighbourhood scale in Glasgow. It will also assess the accuracy of these predictions, along with the best communication method for future users. Chapter 9 will determine the model accuracy in other geographical locations.

8.1. Background

Different types of machine learning methods were considered to predict soil Pb concentrations in Glasgow; however, random forests (RF) were selected. Random forests were chosen because they can quantify uncertainty relatively easily, they are highly flexible i.e. they can be applied to both continuous and categorical datasets, they can model complex linear or non-linear relationships in large datasets, and they are data-driven so require relatively little tuning, which reduces user input and error (Cutler et al., 2007; Hengl et al., 2018). Crucially, RFs produce highly accurate predictions, when compared with other types of machine learning and predictive modelling methods e.g. artificial neural networks (ANN), support vector machines (SVM), and linear mixed models (LMM) (Vega et al., 2010; Chen et al., 2019; Zaorska et al., 2019). For example, Zhang et al. (2020) found validation results from RF and SVM were slightly better, compared with ANN, when predicting soil Pb in Hefei City, China. Similarly, Lacarce et al. (2012) found classification and regression trees (a similar method to RF) outperformed LMM when mapping soil Pb throughout France. Wang et al. (2020) also proved that RFs performed better than land use regression models, when predicting soil Pb in a rural area of China.
8.1.1. Introduction to Random Forests

Random forests use decision-based trees to build a prediction algorithm. Since RFs were introduced over 20 years ago (Breiman, 2001), they have become a popular and widely used machine learning method. Random forests have been implemented in numerous fields, such as soil science (e.g. Gomes et al., 2019; Wang et al., 2020), air pollution (e.g. Chen et al., 2020), hydrology (e.g. Liang et al., 2018; Tyralis et al., 2019), ecology (e.g. Cutler et al., 2007; Li et al., 2017), agriculture (e.g. Jeong et al., 2016; Liakos et al., 2018), remote sensing (e.g. Rodriguez-Galiano et al., 2011), bioinformatics (e.g. Boulesteix et al., 2012), and genetics (e.g. Dettling, 2004; Zaorska et al., 2019).

The main downfall of RFs is that, unlike model-based geostatistics, RFs do not account for spatial autocorrelation between samples. Subsequently, bias may be introduced if sampling was non-random or clustered (Hengl et al., 2018). There have been several proposed solutions to this issue e.g. random forests for spatial prediction (Hengl et al., 2018) or geographical random forests (Georganos et al., 2019). However, these approaches often lead to unrealistic features or artefacts appearing in prediction maps (Hengl et al., 2018). Another disadvantage of RFs is that their complexity can mean that it is harder to interpret drivers of variation.

Simply, RFs operate by combining numerous classification or regression trees (CART) (Figure 8.1). Classification trees are used when the target variable is categorical e.g. high, medium, low soil Pb, whereas regression trees predict continuous variables e.g. a specific concentration of soil Pb (Cutler et al., 2011). Both classification and regression trees are created using binary recursive partitioning to allocate observations to a ‘leaf’ or terminal node. Binary recursive partitioning in regression trees works by splitting the data at each node into two subsets, based on which value or classes of a predictor variable will generate the minimum mean squared residual (Figure 8.1; Cutler et al., 2011). This process is repeated until a terminal node is reached. The predicted value of the target variable, \( \hat{h}(x) \), is the average of the values within the terminal nodes, \( k \), where \( y_{k_1}, \ldots, y_{k_n} \) are the calculated values of soil Pb at terminal nodes in the training data,

\[
\hat{h}(x) = \bar{y}_k = \frac{1}{n} \sum_{i=1}^{n} y_{k_i}.
\]  

(8.1)
**Figure 8.1**: Example of a regression tree built using binary recursive partitioning, with tree structure explained in green, splitting criteria in blue, and soil Pb concentrations at the terminal nodes in red. Own figure made using information in Cutler et al., 2011.
A similar process is used for classification trees, but terminal nodes consist of a class not a value. In both classification and regression trees, tree size is controlled by tuning parameters such as the maximum number of splits or terminal nodes (Figure 8.1; Cutler et al., 2011).

Random forests introduce randomness to the classification or regression tree by testing a new random subset of predictor variables at each node (Cutler et al., 2011). They also create randomness by ‘bagging’ each tree, i.e. each individual tree or ‘base learner’ is built from a bootstrapped sample, with replacement, of the original data (Cutler et al., 2011). The fraction of randomly selected input data used to construct each tree can be altered and is known as the ‘In Bag Fraction’. Samples not used in tree building are thus called ‘Out of Bag data’. Out of bag (OOB) samples are a useful cross-validation tool, as OOB values can be predicted in each tree using the prediction algorithm generated from the In Bag samples. Subsequently, the difference between the actual and predicted values is summed to give the total error (Cutler et al., 2011).

The OOB error may be reduced by optimising the tuning parameters. These consist of the total number of trees in the forest ($J$) - which should be a large enough number for the OOB error to stabilise and odd to prevent ties; the number of predictor variables to sample at each split ($m$) - the default is the total number of variables divided by three; the In Bag fraction, and size of tree or maximum number of splits ($m_{ns}$) - which should be small enough to prevent overfitting (Cutler et al., 2011). Although these tuning parameters can be defined by the user, previous research has shown that little tuning is needed as different tuning values make minimal difference to the OOB error (Lopes, 2015).

Overall, RFs work by bootstrapping a random subset of the original data. Binary recursive partitioning is then used to build a classification or regression tree, where a random subset of predictor variables ($p$) are selected at each splitting point. Tree building is repeated many times, with results from each tree or base learner averaged into one predictive algorithm or strong learner. This reduces over-fitting and generates more accurate results than using a single classification or regression tree. In other words, the RF algorithm can predict the target variable at an unknown point, $\hat{f}(x)$, by averaging each regression tree, $\hat{h}_j(x)$ from Equation 8.1, for $J$ trees (Cutler et al., 2011),

$$\hat{f}(x) = \frac{1}{J} \sum_{j=1}^{J} \hat{h}_j(x).$$  \hspace{1cm} (8.2)
Quantile regression forests (QRFs) (Meinhausen, 2006) are an expansion of RFs which quantify uncertainty in predicted regression values. As shown in Equation 8.2, RF predictions are usually the average prediction for all $J$ trees. However, QRF examines the distribution of all predicted values from the individual regression trees so that QRF predictions can include many different percentiles (Kirkwood et al., 2016). This allows uncertainty to be represented. For example, the 0.05 and 0.95 percentiles would give 90% upper and lower confidence intervals around the median predicted value (Hengl et al., 2018).

### 8.1.2. Studies using random forests to predict soil Pb

Other studies have used RF or QRF to predict soil Pb concentrations in urban and rural environments (Table 8.1), and their models are relatively accurate according to cross-validation statistics (i.e. $R^2$ values $> 0.8$, Table 8.1). However, the studies with the highest $R^2$ values tend to have the lowest soil Pb levels e.g. maximum soil Pb concentration of 34 mg kg$^{-1}$ (Tan et al., 2019) and 81 mg kg$^{-1}$ (Wang et al., 2020), because they do not have to predict Pb hotspots. Examination of all studies using RF to predict soil Pb indicates that there is scope for development.

Firstly, there has been a growing popularity for combining remotely sensed and environmental data as predictor variables (e.g. Kirkwood et al., 2016; Omondi and Boitt, 2020). However, as proven by the poor cross-validation statistics from Kirkwood et al.’s (2016) model, soil Pb is more influenced by anthropogenic factors, and less so by geographical controls (unless they correlate with OM content).

Secondly, model validation can be performed by partitioning samples into training and validation datasets, with the training dataset used to build the model and the validation dataset used to determine model accuracy (e.g. Radocaj et al., 2020, training dataset = 75% of samples, validation = 25%). This is similar to the OOB error but reduces the number of samples available for model building, and limits the size of the validation dataset. Instead, 10-fold cross-validation could be used for model validation, as was done in Kirkwood et al. (2016). 10-fold cross validation assigns each sample to one of ten folds, builds the RF model without samples from a specific fold, then uses the model to predict fold sample values, and calculates the difference between predicted and actual values (Marchant, 2018). This is repeated for each fold, with the results averaged and RMSE (or a similar metric) calculated.
<table>
<thead>
<tr>
<th>Study</th>
<th>Target variable(s)</th>
<th>Predictor variables (covariates)</th>
<th>Method</th>
<th>Validation statistics</th>
<th>Sampling</th>
<th>Pb (mg kg⁻¹)</th>
</tr>
</thead>
</table>
| Lacarce et al., 2012 | Soil Pb (log)      | - pH  
- Soil texture  
- Parent material  
- Land use  
- Population density                                      | CART & LMM              | LOOOCV: RMSE = 7.57  
Mean SSE = 1  
Median SSE = 0.43 | Mainland France  
Mixed  
16km²  
NR  
NR                        |              |
| Schwarz et al., 2013    | Soil Pb            | - Housing age  
- Road distance  
- Building distance                                                   | RF - regression         | LOOOCV: R² = 0.42  
External validation: overall accuracy = 72% | Baltimore (USA)  
Urban  
0.08m²  
363  
7 - 9,151*                       |              |
| Wang et al., 2015      | Soil heavy metals  
(Pb, and cadmium (Cd)) | - Background value  
- Atmospheric and water values  
- Urban sources (e.g. pop density)  
- Agrl. Sources (e.g. fertiliser use)  
- Industrial sources (dist. to plant) | RF - regression         | 10-fold cross-validation:  
AUC = 0.86  
(Good)  
R² = 0.74 | Guangdong Province (China)  
Agricultural  
250  
Local  
233  
10 - 4,770                      |              |
| Kirkwood et al., 2016  | Soil elements (including Pb), pH and OM                                                                | Geophysical data (e.g. elevation & slope, n=6)  
Remote sensed data (e.g. landst spectral indices n=20) | QRF                     | 10-fold cross-validation:  
RMSE = 43  
R² = 0.14 | SW England  
Rural  
568  
12.2km²  
49  
23 - 201                      |              |
| Shaheen & Iqbal, 2018  | Soil heavy metals  
(Pb, Cd, chromium (Cr)) | - Soil depth (n=3)  
- Physio-chemical parameters (e.g. OM, pH, Cd, n=8) |
| Hong et al., 2019       | Soil heavy metals (Pb, Zn) | - Spectral data from field samples (n=2017)  
- | RF - regression         | Split validation set:  
RMSE = 3.73  
R² = 0.82 | Wuhan (China)  
Peri-urban agricultural  
170  
ND  
~2.9km²  
35.91  
23 - 62                      |              |
<table>
<thead>
<tr>
<th>Study</th>
<th>Target variable(s)</th>
<th>Predictor variables (covariates)</th>
<th>Method</th>
<th>Validation statistics</th>
<th>Sampling</th>
<th>Pb (mg kg⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tan et al., 2019</td>
<td>Soil heavy metals (Pb, Zn, Cd, arsenic (As))</td>
<td>- Spectral data from field samples (n=33)</td>
<td>RF - regression</td>
<td>Split validation set: RMSE = 1.2  R² = 0.93</td>
<td>Jiangsu (China) Mining area</td>
<td>30</td>
</tr>
</tbody>
</table>
| Wang et al., 2020  | Soil heavy metals (including Pb)    | - Population density  
- Distance to sewage river, res. area, road, vacant land (n=8) | RF - regression | LOOCV: RMSE = 3.59  R² = 0.897 | Dongli (China) Rural | 393 | ND ~1.2km² | 33 | 17.81 |
| Radocaj et al., 2020 | Soil heavy metals (Pb, Cd, Cr, copper (Cu), nickel (Ni), Zn) | - Land cover, DEM, and slope (n=3)  
- Soil pH, texture, C:N ratio and CaCO₃ content (n=6) | RF - regression | Split validation set: RMSE = 4.9  R² = 0.46 | Mainland Croatia Mixed | 469 | ND ~66km² | 33 | 0.405 |
| Omondi & Boiitt 2020 | Soil heavy metals (Pb, Zn, Cd) | - Hyperspectral remotely sensed data (spectral indices)  
- Environmental parameters (e.g. slope, and distance to hotspots e.g. industries and roads, n=30) | RF - regression | OOB error: RMSE = 0.44  R² = 0.83 | Nairobi River (Kenya) Rural | 60 | ND ~0.25km² | 217 | 7.457 |
| Zhang et al., 2020  | Soil heavy metals (Pb, Zn, Cu, Cd, mercury (Hg)) | - Soil properties  
- Land use type  
- Urbanisation history | RF (and others) | Split validation set: RMSE = 7.66  R² = 0.79 | Hefei, City, E. (China) Urban | 251 | 2 km² | 29.4 | 17.91 |
| Wragg and Cave, 2021 | Soil heavy metals and bioaccessibility (Pb, and As) | - Geology (n=2)  
- Historic industry locations (IDW distance to, n=5) | RF | OOB data NR | Stoke-on-Trent (UK) Urban | 747 | 250 m | NR | 16-1,234 |
RF predictions could also be developed by testing the prediction models on different locations with an independent validation dataset. This checks that models have not been overfit to local conditions, which is more likely with the more covariates included. Consequently, the number of covariates included in RF model is often reduced to prevent overfitting. This is achieved by first running the RF with excess variables (i.e. >20), calculating the predictor importance values to highlight the most influential variables, and then re-running the RF with fewer variables (e.g. Wang et al., 2020). This generates good cross-validation statistics but may be at risk of data mining or ‘p hacking’ (Lark 2017). Therefore, appropriate covariates should be carefully selected before including a variable in the RF model.

Another potential improvement is including error margins or confidence intervals alongside predictions to highlight that predicted values are not 100% accurate. This could be achieved using QRF, as shown by Kirkwood et al. (2016). Alternatively, some measure of risk could be quantified from the predicted values.

From a risk management perspective, it is more important to correctly predict ‘high’ soil Pb values, as these soils pose a greater potential risk. However, urban soil geochemical datasets tend to contain mostly low concentrations from diffuse background pollution (e.g. Matschullat et al., 2000; Rawlins et al., 2005); hence, model accuracy is predominantly a measure of how well these low values are estimated. Soils with higher metal concentrations are less common as they form via deposition or emissions from a specific source or point, which may be unique to a region (Rawlins et al., 2005). Overall, there tends to be less data on soil hotspots, and they are more difficult to predict, meaning few studies have attempted predictive soil Pb hotspot modelling using RFs.

One study which did attempt soil Pb hotspot modelling is Schwarz et al. (2013). It is unique as their RF model predicts soil Pb in an exclusively urban environment (Baltimore); therefore, there is a greater range of soil Pb concentrations to estimate (Table 8.1). The study is also unusual as they only use three covariates, all of which are proven to correlate with soil Pb. They also determine the accuracy of predicting high soil Pb concentrations, by classifying the RF predictions as high (>400 mg kg\(^{-1}\)) or low (<400 mg kg\(^{-1}\)). External validation indicates that the total accuracy of their RF model is 72% (Schwarz et al., 2013). However, their external validation dataset contained only 29 samples, two of which had soil Pb >400 mg kg\(^{-1}\).
8.1.3. Defining high soil Pb values

If a model is used to identify high soil Pb levels, then the next logical question is what constitutes high soil Pb concentrations? This topic was explored in Section 2.3. (How much soil Pb is too much?) but will be summarised here.

Recent UK government guidance has moved away from soil guideline values (SGVs) to use Category 4 Screening Levels (C4SLs) in order to assess if soil is at risk from contamination (DEFRA, 2014). The C4SLs were derived using Contaminated Land Exposure Assessment (CLEA) deterministic exposure models, which combines environmental and toxicological factors (DEFRA, 2014). These generate generic risk assessment soil Pb concentrations for six different land uses, below which land is not considered to be contaminated (Table 8.2; DEFRA, 2014). Probabilistic modelling was also applied in order to determine that a random human receptor would not exceed a concerning blood Pb level for a given soil Pb concentration. This indicated that the C4SLs are conservative (DEFRA, 2014).

<table>
<thead>
<tr>
<th>Land use</th>
<th>C4SL (Pb mg kg⁻¹)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allotments</td>
<td>80</td>
</tr>
<tr>
<td>Residential (with home-grown produce)</td>
<td>200</td>
</tr>
<tr>
<td>Residential (no home-grown produce)</td>
<td>310</td>
</tr>
<tr>
<td>Residential Open Space</td>
<td>630</td>
</tr>
<tr>
<td>Public Open Space (e.g. park)</td>
<td>1300</td>
</tr>
<tr>
<td>Commercial</td>
<td>2300</td>
</tr>
</tbody>
</table>

For the purposes of this study any soil with $\geq 200$ mg kg⁻¹ of soil Pb will be classified as ‘high’. The value was selected from the residential/garden C4SLs guideline values for soil Pb (Table 8.2), as residential gardens are widespread in cities, thus more people are likely to come into contact with garden soil compared with allotment soil. It is important to emphasise that this definition of high soil Pb covers soil that could potentially cause risk to human health, as the C4SL values are intentionally cautious (DEFRA, 2014). Therefore, all sites exceeding this value for this land use type should be investigated in more detail i.e. run a source-pathway-receptor risk assessment, before any firm conclusions are drawn.

The ‘high’ soil Pb threshold could be adjusted if other land uses were of particular interest e.g. allotments, or if the decision maker was risk-adverse.
8.2. **Chapter aims**

Given the previous literature and overall aims of this project, the two key questions to be addressed in this chapter are:

1. Can a random forest model accurately predict soil Pb concentrations and/or identify urban locations at risk from high soil Pb concentrations at the **city** scale?

2. Can a random forest model accurately predict soil Pb concentrations and/or identify urban locations at risk from high soil Pb concentrations at the **neighbourhood** scale?

Each question can be broken down into two parts: a) predictions of soil Pb concentrations; b) identification of risk from high soil Pb concentrations. Furthermore, consideration will be given throughout on the most effective presentation methods to communicate the results to decision makers.

8.3. **Methodology**

The methodology used to generate regression RF and QRF models was loosely based on Kirkwood *et al.* (2016), except Matlab was used to run the code instead of R. Specifically, the Glasgow training dataset was first loaded into Matlab, which included measured soil Pb concentrations and values of predictor variables/covariates. Covariates were selected, transformed, and classified, based on previous geostatistical modelling (see Sections 8.4.1.1. and 8.4.1.2.). Mean soil Pb concentrations from each field site were log transformed due to their high skew (skew in Glasgow G-BASE=18.33, Bishopbriggs=2.96, Paisley=1.79) and set as the target variable.

The selected covariates at each sample point were then determined. This was achieved in QGIS® or ArcGIS® by digitisation and nearest neighbour joins (or field notes where no dataset was available, Table 8.3). This follows the previous covariate determination methodology described in Section 6.3.1.
Table 8.3: Details on determining covariate values at grid or sample points for the Glasgow G-BASE dataset. Blue = continuous covariate. Orange = categorical covariate. GIS includes QGIS® and ArcGIS®.

<table>
<thead>
<tr>
<th>Covariate</th>
<th>Data source</th>
<th>Methodology</th>
<th>Groups/Transform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Organic matter (OM) %</td>
<td>Loss on ignition (LOI) during analysis (Johnston et al., 2005)</td>
<td>For grid points OM was simply kriged using known values at sample points to give a full coverage of the study area. Spatial join between prediction grid and kriged grid points.</td>
<td>Log transform</td>
</tr>
<tr>
<td>Building</td>
<td>Digitised historic maps from Digimap (Crown Copyright, 2021a)</td>
<td>In GIS digitised any building footprints (including gardens) from a historic map dated close to 1920, added any new buildings on the historic map from ~1970, and on a current map. This gives the first date of building not the current building age. Spatial join with grid/sample points to get distance to and closest building polygon. Any point &gt;5 m from a building polygon was classed as 0 i.e. remote from buildings. Reclassed 1920 to pre-1920, 1970 to 1920-1970, and current map to post-1970.</td>
<td>Pre-1920 1920-1970 Post-1970 0</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td>Log transform</td>
<td></td>
</tr>
<tr>
<td>Distance</td>
<td>Digitised historic maps from Digimap (Crown Copyright, 2021a)</td>
<td>In GIS digitised footprint of any industrial sites from historic maps in study area (plus 10 m buffer). Largest building footprint used. Date of appearance on first map observed on and type of industry recorded as attributes. Spatial join with grid/sample points to get distance to and attributes of the nearest historical industry polygon. Any point &gt;40 m from an industry polygon was classed as 0 (age) or non-industrial (NHI type). Light (LHI) = raw materials and transport industries. Heavy (HHI) = metal, chemical, and waste industries.</td>
<td>Pre-1920 1920-1970 Post-1970 0</td>
</tr>
<tr>
<td>Age</td>
<td></td>
<td>LHI (light) HHI (heavy) NHI (no)</td>
<td></td>
</tr>
<tr>
<td>Type</td>
<td></td>
<td>Log transform</td>
<td></td>
</tr>
<tr>
<td>Dist.</td>
<td>Kernel density estimation (KDE) of historic industry centroids</td>
<td>In GIS determined historic industry centroids with their coordinates (centroid tool then add geometry info). Loaded centroids into Crimestat (Levine, 2010) as a primary file, run a single KDE with a fixed interval of 500 m and normal interpretation to determine the absolute density per km² for each sample/grid point. Saved as .dat file, loaded into excel and reformatted as .csv, finally load into GIS.</td>
<td>Log transform</td>
</tr>
<tr>
<td>Density</td>
<td></td>
<td>Log transform</td>
<td></td>
</tr>
<tr>
<td>Road distance</td>
<td>OS MasterMap road polylines (Crown Copyright, 2021b)</td>
<td>In GIS, clipped roads to study area (plus 10 m buffer). Added 3 m buffer to road polylines. Nearest neighbour join determined distance to nearest road plus buffer for each grid/sample point.</td>
<td>Log transform</td>
</tr>
<tr>
<td>Land use</td>
<td>OS Greenspace layer (Crown Copyright, 2021b)</td>
<td>In GIS merged all Greenspace layers covering study area. Clipped to study area extent (plus 10m buffer). NN join to determine nearest greenspace polygon to grid/sample point. Reclassified in Matlab, road verge=amenity-transport; garden=private garden or allotments; commercial=sports facility, tennis court, bowling green, civic space, religious grounds, institutional grounds, camping/caravan park, golf course, cemetery or play space; residential open space=amenity-residential or business, open space=public park or garden. Other classification schemes were explored (see Chapter 6).</td>
<td>- Natural</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Road verge</td>
<td>- Garden</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Commercial</td>
<td>- Res. open space</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Open space</td>
<td></td>
</tr>
</tbody>
</table>
The Treebagger function (MathWorks®, 2021a) was used to build the RF algorithm, and as the soil Pb measurements are a continuous variable, regression trees were used as base learners (example code provided in Appendix 8.1). Other parameters in the treebagger algorithm can be adjusted to avoid over-fitting to outliers. These were optimised after experimentation with the OOB error, and include the in-bag fraction, which was set to 0.5; the maximum number of splits, which were controlled at six; and number of randomly selected predictor variables tested at each split point, which was set at four (the default 1/3 of predictor variables was too low due to fewer predictor variables than the average RF). The total number of trees in the forest \((J)\) was set at 1001, as the odd number prevents possible ties, and the number is large enough for the OOB error to stabilise at its minimum level (which was assessed by plotting OOB vs number of trees).

The predictor importance (MathWorks®, 2021b) is calculated from the OOB error for each covariate, and was also output from the RF. This indicates which predictor variables have most relative influence on soil Pb concentrations. Furthermore, the partial dependence plots (PDP; MathWorks®, 2021c) for each covariate were output. They depict the average partial relationship between soil Pb (log) and each covariate value or class from each trained regression tree (MathWorks®, 2021c). These plots provide a more detailed picture of the relationship between soil Pb and each covariate.

Quantile regression forests used the RF algorithm to predict soil Pb values for a range of percentiles (0.05, 0.25, 0.4, 0.5, 0.6, 0.75, 0.95). Predicted soil Pb (log) concentrations for the 5\(^{th}\), 50\(^{th}\), and 95\(^{th}\) percentiles were output and mapped in ArcGIS® to determine median soil Pb predictions and 90% confidence intervals.

The models were validated using 10-fold cross validation for a range of prediction percentiles. This operated by randomly assigning each measurement to one of 10 folds, the RF model was then re-built without values from this fold, and fold soil Pb values were predicted using QRF for a range of percentiles. This was repeated for each fold and the predicted values are averaged for each percentile. For each percentile the predicted soil Pb values \((\mathbf{P}_i)\) were compared with the observed soil Pb values \((\mathbf{O}_i)\). This was done visually in a scatterplot, with a 1:1 line indicating accurate predictions, and values above the line representing over-prediction and values below the line underprediction. Quantitative assessments were also
performed by calculating the average Mean Error \((P_i - O_i)\), average Root Mean Square Error (RMSE):

\[
RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (P_i - O_i)^2}
\]

Pearson’s correlation coefficient \((p)\), and the concordance correlation coefficient \((\hat{p}_c)\) or CCC; Lawrence and Lin, 1989). The CCC \((\hat{p}_c)\) is a more robust measure of correlation, as it combines Pearson’s correlation coefficient, \(p\), with the variance of the predicted and observed values, \(\sigma_{P/O}\), and their mean values, \(\mu_{P/O}\) (Lawrence and Lin, 1989):

\[
\hat{p}_c = \frac{2p\sqrt{\sigma_p\sigma_o}}{\sigma_p + \sigma_o + (\mu_p - \mu_o)^2}
\]

This assessed the divergence of a pair of observations from the 1:1 line, whereas the Pearson’s correlation coefficient measured the deviance from the best fit line. As with Pearson’s, an ideal coefficient will be close to 1. This validation process was also repeated for only high-Pb samples (soil Pb ≥200 mg kg\(^{-1}\)) and low-Pb samples (soil Pb <200 mg kg\(^{-1}\)), in order to assess model accuracy for different soil Pb concentrations.

To determine the risk of high soil Pb concentrations QRF were used. Quantile regression forests used the RF model to predict soil Pb values for every percentile from 0.01 to 0.99. A threshold soil Pb level was then defined (soil Pb log = 5.23, = 200 mg kg\(^{-1}\)). The percentile at which the predicted soil Pb value exceeded this threshold was determined for each point. This percentile was multiplied by 100 to give its probability (%) of having high soil Pb. These probabilities were extracted and mapped in ArcGIS® for Greater Glasgow, Paisley, and Bishopbriggs. Furthermore, a Getis-Ord Gi* (Getis & Ord, 1992) hotspot analysis using inverse distance weighting was run to identify significant (>95% confidence) clustering of probability values. The hotspots were shown as contour lines on the final map to help pinpoint the most at-risk locations from high soil Pb.

The same process was repeated for a 200 m by 200 m prediction grid covering Greater Glasgow and a 5 m by 5 m prediction grid over the Paisley and Bishopbriggs study areas. The prediction grid was previously generated in ArcGIS® and combines grid point coordinates with the value for each covariate, determined using a nearest neighbour join (Table 8.3). As there was no organic matter data covering the entire area sampled OM values were kriged to match the grid area. The covariate values were carefully formatted to match the input data. Quantile
regression forests for every percentile was then run, and results were compared with the threshold. The probability of each grid point having high soil Pb was extracted and plotted in ArcGIS®, alongside predicted soil Pb values.

The predicted probability values were validated by comparison with observed soil Pb values. Observed sample values were classified as high if they contained soil Pb ≥200 mg kg\(^{-1}\) and low if soil Pb was <200 mg kg\(^{-1}\). The predicted probabilities were classified as high-risk if their probability was greater than or equal to a cut-off value (\(C\%\)) and low risk if their probability was <\(C\%). This cut-off value, \(C\%\), was decided by receiver operating characteristic curves (ROC) which compares the true positive and false positive classification rates for high-risk samples at different probabilities or cut-off values (Fawcett, 2006). The optimal cut off value for \(C\) was the maximum Youden’s index \((YI; \text{Youden}, 1950)\). The Youden’s index finds the best balance between the true positive rate \((TPR)\) i.e. number of samples correctly categorised in a group divided by total number of samples in the group) and the false positive rate \((FPR)\) i.e. number of samples incorrectly categorised in a group divided by total number of negative samples/samples not in the group):

\[
YI = TPR + FPR - 1.
\] (8.5)

Subsequently, a confusion matrix compared the predicted and observed classes to assess risk prediction accuracy. The full methodology is summarised in Figure 8.2, and was repeated for Paisley, Bishopbriggs, and all of Greater Glasgow (G-BASE) samples.
Figure 8.2: Flow diagram summarising the methodology for building and validating a RF model. The diagram also shows how QRF predicts soil Pb values and the probability of a sample having high soil Pb contents. For Glasgow, Paisley, and Bishopbriggs.
8.4. **Results and discussion**

8.4.1. **City scale predictions**

This section aims to address the first research question, i.e. can a RF model accurately predict soil Pb concentrations and/or risk at the city scale?

8.4.1.1. **Covariate selection**

Six covariates were selected for the Greater Glasgow G-BASE RF model based on previous geostatistical analysis. This number was selected with a view to maximising model accuracy while limiting overfitting. The selected covariates were OM percentage (log), building age, historic industry density (log)/age/type, and land use (shaded green in Table 8.4), because the relationship between soil Pb and these six covariates is strong and agrees with the literature. The two covariates shaded in blue (Table 8.4) - soil texture and road type - could have been included in the RF model as they showed a significant relationship with soil Pb in the geostatistical analysis. However, further investigation indicated that the relationship between soil Pb and these covariates was dominated by a few outliers was not representative of relationships in the general dataset. Including these covariates in the RF model would have increased the risk of over-fitting the model to only represent the conditions in Glasgow.

Overall, the predictor importance plot (Figure 8.3) indicates that a broadly similar relationship is found between soil Pb and the covariates in the RF model, compared with the geostatistical analysis (Table 8.4). The general similarity between the methods justifies including each predictor in the Glasgow RF model.
8.4.1.2. Prediction of soil Pb concentrations

The RF model was used with QRF to predict soil Pb concentrations at each G-BASE point in Greater Glasgow, but also at a 200 by 200 m prediction grid covering the study area. For the prediction grid, the median soil Pb (log) estimations were first classified into six equal intervals (Figure 8.4). The predicted values were also classified by the most precautionary land use C4SL exceeded at each point, regardless of the land use at each point (Table 8.2; Figure 8.5A). In order to communicate uncertainty QRF was used to predict the 5th and 95th percentiles to give a 95% confidence interval (Figure 8.5B and C). In all maps the classes were coloured according to the colorbrewer scheme (Brewer, 2006) which is designed to help the user easily differentiate between different classes. Furthermore, higher values were given a darker hue, as studies have observed people associated darker hues with more risk and vice versa (MacEachren, et al., 2005).

Figure 8.3: Predictor importance plot from the RF model on Glasgow G-BASE dataset. Covariates are ordered from strongest relationship with soil Pb on the left to weakest relationship on the right (based on geostatistical analysis - see Chapter 6).
Figure 8.4: Map of Greater Glasgow (200 by 200 m prediction grid) showing median estimated soil Pb concentrations (log) predicted from a QRF algorithm, classified into six equal intervals.
Figure 8.5: Maps of Greater Glasgow showing (A) the median predicted soil Pb concentration, (B) the lower confidence interval/5th quantile predictions, and (C) the upper confidence interval/95th quantile predictions. Determined from QRF at prediction grid points (200 by 200 m) covering the study area. Classified by the land use C4SL exceeded at each prediction grid point, regardless of current land use (see Table 8.2 for values). N.B. most precautionary land use C4SL is shown, so if the point is classified as garden, the allotment C4SL is also exceeded.
The first map showing logarithmic values of soil Pb classified into equal intervals (Figure 8.4) highlights the variability in predicted soil Pb concentrations. Hotspots are clear in the east end and in Paisley, and lower values are predicted in the city centre. However, these values are not easy to interpret, as there is no measure of risk included.

Comparatively, when the predicted values are classified by which land use C4SL are exceeded (Figure 8.5) the variability in the dataset is less evident, but this map is perhaps better at communicating risk. Specifically, the maps in Figure 8.5 show that only one to three C4SLs are likely to be exceeded, depending on which percentile is selected. For instance, median soil Pb in Greater Glasgow is likely to exceed only the allotment C4SL i.e. fall between 80 and 200 mg kg\(^{-1}\). This homogeneity is partly due to the central tendency bias that has been associated with RF predictions. Specifically, RFs use weighted averages when combing multiple trees, and this causes predicted values to be brought closer to the mean (Kirkwood et al. 2016; Zhang and Lu, 2012).

The upper and lower confidence intervals maps (Figure 8.5B and C) are presented beside the median value prediction map to show uncertainty in the predicted median values. Side by side or static presentation allows the user to compare how the predicted values, and the uncertainties in those predictions, vary from cell to cell, with ‘data’ on the left and ‘uncertainty’ on the right of the figure (MacEachren et al., 2005). This type of display was favoured by decision makers when communicating uncertainty in urban planning issues, according to a web-survey conducted by Aerts et al. (2003).

The uncertainty maps indicate extreme Pb values that the prediction is unlikely to fall out with, so have many practical uses (see Section 8.4.3.). The choice of which percentile is mapped is also adaptable. In this case, the 95\(^{th}\) percentile was selected to represent the upper confidence interval, but a risk adverse decision maker might want the 99\(^{th}\), or even 99.9\(^{th}\) percentile, mapped instead.

The current maps (Figure 8.5) consider any future land use rather than the current land use at any given location. Namely, they highlight areas where it could be problematic to develop allotments as Pb concentrations are likely to exceed the C4SL for allotment soil. However, the C4SL maps could be expanded further by combining them with information on current land use at each grid point to show
areas where soil Pb concentrations currently pose a potential risk to human health and should be investigated in more detail.

The spatial pattern of the median and 95th percentile predictions do not always match due to differences in individual regression trees. Therefore, the difference between these predictions was calculated and mapped (Figure 8.6). This highlights which areas of Glasgow are most uncertain with regard to soil Pb predictions i.e. the City Centre and West Glasgow (around Paisley). The most uncertain locations tend to correlate with the highest density of historic industry. This suggests that the covariate values at these locations can cause a large range of soil Pb concentrations.

**Figure 8.6:** Difference between 95th percentile and median predictions for QRF estimations at the Greater Glasgow prediction grid.
The accuracy of the Glasgow G-BASE model was assessed via 10-fold cross validation (Table 8.5). Validation statistics were calculated for different soil Pb groups (all samples and just high or low samples), and for a range of percentiles (0.05, 0.25, 0.4, 0.5, 0.6, 0.75, 0.95). The predicted and observed soil Pb (log) values were also plotted by colour (high=pink, low=blue), and compared with a 1:1 line to give an indication if values had been under or over-predicted (Figure 8.7).

<table>
<thead>
<tr>
<th>Soil Pb group</th>
<th>Prediction percentile</th>
<th>Mean ME</th>
<th>Mean RMSE</th>
<th>CCC</th>
<th>Pearson R²</th>
<th>Soil Pb (log) range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual</td>
<td>n/a</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>2.8 - 9.2</td>
</tr>
<tr>
<td>5ᵗʰ</td>
<td>-0.95</td>
<td>0.96</td>
<td>0.1</td>
<td>0.42</td>
<td></td>
<td>2.93 - 4.75</td>
</tr>
<tr>
<td>25ᵗʰ</td>
<td>-0.39</td>
<td>0.53</td>
<td>0.25</td>
<td>0.5</td>
<td></td>
<td>3.78 - 5.35</td>
</tr>
<tr>
<td>40ᵗʰ</td>
<td>-0.17</td>
<td>0.45</td>
<td>0.3</td>
<td>0.51</td>
<td></td>
<td>3.98 - 5.58</td>
</tr>
<tr>
<td>50ᵗʰ</td>
<td>-0.05</td>
<td>0.44</td>
<td>0.33</td>
<td>0.51</td>
<td></td>
<td>4.18 - 5.78</td>
</tr>
<tr>
<td>60ᵗʰ</td>
<td>0.09</td>
<td>0.45</td>
<td>0.34</td>
<td>0.52</td>
<td></td>
<td>4.47 - 5.93</td>
</tr>
<tr>
<td>75ᵗʰ</td>
<td>0.35</td>
<td>0.53</td>
<td>0.33</td>
<td>0.53</td>
<td></td>
<td>4.67 - 6.17</td>
</tr>
<tr>
<td>95ᵗʰ</td>
<td>1.09</td>
<td>1.1</td>
<td>0.17</td>
<td>0.5</td>
<td></td>
<td>5.17 - 8.75</td>
</tr>
<tr>
<td>All n=1,285</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5ᵗʰ</td>
<td>-1.71</td>
<td>1.71</td>
<td>0.01</td>
<td>0.07</td>
<td></td>
<td>3.31 - 4.75</td>
</tr>
<tr>
<td>25ᵗʰ</td>
<td>-1.1</td>
<td>1.1</td>
<td>0.01</td>
<td>0.07</td>
<td></td>
<td>3.91 - 5.32</td>
</tr>
<tr>
<td>40ᵗʰ</td>
<td>-0.87</td>
<td>0.87</td>
<td>0.02</td>
<td>0.09</td>
<td></td>
<td>4.18 - 5.58</td>
</tr>
<tr>
<td>50ᵗʰ</td>
<td>-0.73</td>
<td>0.73</td>
<td>0.04</td>
<td>0.12</td>
<td></td>
<td>4.47 - 5.78</td>
</tr>
<tr>
<td>60ᵗʰ</td>
<td>-0.58</td>
<td>0.6</td>
<td>0.06</td>
<td>0.16</td>
<td></td>
<td>4.61 - 5.93</td>
</tr>
<tr>
<td>75ᵗʰ</td>
<td>-0.27</td>
<td>0.39</td>
<td>0.2</td>
<td>0.28</td>
<td></td>
<td>4.83 - 6.17</td>
</tr>
<tr>
<td>95ᵗʰ</td>
<td>0.49</td>
<td>0.52</td>
<td>0.41</td>
<td>0.62</td>
<td></td>
<td>5.39 - 8.75</td>
</tr>
<tr>
<td>High (≥5.3 log Pb or ≥200 mg kg⁻¹) n=351</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5ᵗʰ</td>
<td>-0.67</td>
<td>0.67</td>
<td>0.14</td>
<td>0.48</td>
<td></td>
<td>2.9 - 4.51</td>
</tr>
<tr>
<td>25ᵗʰ</td>
<td>-0.13</td>
<td>0.31</td>
<td>0.36</td>
<td>0.48</td>
<td></td>
<td>3.78 - 5.11</td>
</tr>
<tr>
<td>40ᵗʰ</td>
<td>0.09</td>
<td>0.29</td>
<td>0.33</td>
<td>0.43</td>
<td></td>
<td>3.98 - 5.33</td>
</tr>
<tr>
<td>50ᵗʰ</td>
<td>0.21</td>
<td>0.33</td>
<td>0.27</td>
<td>0.39</td>
<td></td>
<td>4.18 - 5.53</td>
</tr>
<tr>
<td>60ᵗʰ</td>
<td>0.33</td>
<td>0.39</td>
<td>0.2</td>
<td>0.36</td>
<td></td>
<td>4.47 - 5.77</td>
</tr>
<tr>
<td>75ᵗʰ</td>
<td>0.58</td>
<td>0.59</td>
<td>0.12</td>
<td>0.3</td>
<td></td>
<td>4.67 - 6.11</td>
</tr>
<tr>
<td>95ᵗʰ</td>
<td>1.31</td>
<td>1.31</td>
<td>0.04</td>
<td>0.22</td>
<td></td>
<td>5.17 - 7.42</td>
</tr>
<tr>
<td>Low (&lt;5.3 log Pb or &lt;200 mg kg⁻¹) n=934</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 8.7: Predicted vs actual soil Pb (log) values for Glasgow G-BASE points using the RF and QRF for a range of percentiles (title of each plot). Points are coloured blue if their actual soil Pb is low (<200 mg kg⁻¹) and pink if actual soil Pb is high (≥200 mg kg⁻¹). A 1:1 line has been added to aid interpretation. X and Y axis ranges have been fixed to match.
The Glasgow G-BASE RF model is reasonably accurate at predicting soil Pb values according to 10-fold cross validation. As the median prediction for all samples has a sufficient CCC (0.33), high Pearson’s $R^2$ value (0.51), and mean error (ME) and RMSE close to 0 (-0.05 and 0.44; Table 8.5).

These validation statistics are for the median prediction (i.e. the 50th percentile); however, soil Pb was also predicted for a range of percentiles and compared with observed values (Table 8.5; Figure 8.7). These results suggest that predicted values at the 60th percentile have a slightly closer match with the observed values, as the Concordance and Pearson’s correlation coefficients are slightly higher (0.34 and 0.52 respectively), although the ME and RMSE are slightly worse (0.09 and 0.45 respectively; Table 8.5). This could be due to higher predicted values for a few high outliers at the 60th percentile, thus bringing them closer to the 1:1 line and improving the correlation coefficients (Figure 8.7). However, the more numerous low values will be slightly overpredicted at the 60th percentile, thereby increasing the ME and RMSE.

The validation statistics were also split by soil Pb value; high, and low. Unsurprisingly, the high values are more accurately predicted at higher percentiles (75th and 95th), and the low values at lower percentiles (25th and 40th; Table 8.5). This is also shown visually on the scatter plots in Figure 8.6. Interestingly, if extremely high soil Pb values (>630 mg kg$^{-1}$) are validated, the CCC is 0.97 at the 95th percentile, as the high outliers fall very close to the 1:1 line (Figure 8.7). However, at the 95th percentile the majority of samples have lower soil Pb concentrations which are greatly overpredicted i.e. fall above the 1:1 line (Figure 8.7), resulting in poor validation statistics for the low group and all samples at the 95th percentile (Table 8.5).

In order to efficiently target future sampling and remediation, it is more important for a model to accurately predict high soil Pb concentrations and to err on the side of caution by over-predicting soil concentrations rather than underpredicting. This can be achieved in the RF model by using QRF at high percentiles i.e. the 95th. However, validation statistics are likely to be poor.

The main issue with the RF model is that the predicted values have a much narrower range than the observed values, which is evident when the predicted versus observed values are plotted (e.g. Figure 8.7). This is especially prevalent for the median or 50th percentile predictions (Table 8.5; Figure 8.7) and is likely caused by
a few high outliers in the observed soil Pb dataset. These outliers may have elevated soil Pb concentrations due to point pollution, which is difficult to replicate in the model. Furthermore, the 1,001 regression trees will average out any trees or covariate combinations generating these extreme values. This central tendency bias was discussed earlier and was also observed by Kirkwood et al. (2016) and Zhang and Lu (2012).

This study shows an important step forward in machine learning modelling of urban soil Pb concentrations. Although other research using RF models to estimate soil Pb values have quoted Pearson’s $R^2$ values of 0.93 (Tan et al., 2019) and 0.9 (Wang et al., 2020), they only had to predict Pb concentrations up to a maximum of 81 mg kg$^{-1}$ (Table 8.1) as they studied rural or mining areas. There are no studies which measured the CCC when predicting soil Pb, but if the CCC is compared with the Pearson’s correlation coefficient, which tends to be higher, the results from this study (Pearson’s $R^2 = 0.51$, maximum soil Pb = 9,937 mg kg$^{-1}$) show a better prediction than others with a comparable soil Pb range (e.g. Radocaj et al. (2020) Pearson’s $R^2 = 0.46$, maximum soil Pb = 405 mg kg$^{-1}$, and Schwarz et al. (2013) Pearson’s $R^2 = 0.42$, maximum soil Pb = 9,151 mg kg$^{-1}$; Table 8.1).

Despite these improvements there is scope for development. Perhaps other covariates need to be explored or a different method entirely should be employed. However, it could simply be that urban soil Pb concentrations are too complex to predict with 100% accuracy, using modelling or machine learning techniques. This is because there are so many processes, activities, and land uses that could cause elevated soil Pb levels; therefore, it is not possible to characterise all of these factors within easily obtainable covariates. It is also important to remember that the covariates were carefully selected in order to not overfit the model to the conditions in Glasgow. Including additional covariates would have improved cross-validation statistics but would likely have reduced the model accuracy in other urban regions (see Chapter 9).
8.4.1.3. Prediction of soil Pb risk

Along with soil Pb concentrations, the probability that a location will have high (≥200 mg kg\(^{-1}\)) soil Pb was also calculated for the Greater Glasgow prediction grid. The probabilities ranged from 7-64%, with a mean of 27%. The probability at each grid point was mapped (Figure 8.8) using a similar colour scheme as previous maps. As probabilities are more intuitively understood than actual values, they were simply split into decile classes for mapping. Furthermore, contour lines showing >95% significant hotspots were added.

The probability map shows two major hotspots, specifically one in Paisley and another in the East End (between the city centre and Parkhead; Figure 8.8). The hotspot in the East End likely occurs due to the high density of historic industry (four times higher than the rest of Glasgow) and old housing (97% pre-1920) in this location.

The probability map (Figure 8.8) is more spatially varied and arguably more intuitively understood than the predicted soil Pb maps (Figure 8.5). Probability also communicates uncertainty in a single map; thus, probability maps are widely used in environmental science due to their easy interpretability e.g. to map nitrate groundwater contamination (Sajedi-Hosseini et al., 2018); and earthquake risk (Jena et al., 2020). The probability map could also have classified the soil Pb risk as very low, low, medium, high, and very high, depending on the probability of exceeding a threshold at each grid point.

The main drawback of the probability maps is that they only show high soil Pb locations for garden land uses, whereas Figure 8.5 shows where soil Pb values are likely to exceed the C4SL for multiple land uses. Additionally, the probability methodology is slightly more complex, so needs to be communicated clearly to decision makers.

Overall, the spatial patterns in the probability and C4SL maps (Figures 8.5 and 8.8) tend to agree with each other i.e. there are the same hotspots over the East End (Parkhead) and in Paisley. They also largely agree with measured soil Pb concentrations (see Figure 5.6A). Which map is more useful depends on the end user of the map, which is discussed in more detail in Section 8.4.3.
Figure 8.8: Map of grid points in Greater Glasgow coloured by their probability of having high (≥200 mg kg⁻¹) soil Pb based on QRF predictions on the G-BASE RF model. Contour lines indicate >95% significance of belonging to a high probability hotspot. Locations of interest have also been added.
In order to quantitively assess QRF predictions a high-risk cut value off was determined using a ROC curve and Youden's index (Figure 8.9). The ROC curve is concave and has an under the curve area (AUC) of 0.79, indicating good model predictions (Hastie et al., 2009). The Youden’s index suggests that at least a 30% probability of having high soil Pb gives the best balance between the true positive rate i.e. the number of samples correctly classified as high-risk divided by the total number of samples with ≥200 mg kg\(^{-1}\) of soil Pb, and the false positives rate i.e. the number of samples correctly classified as low risk divided by the total number of samples with <200 mg kg\(^{-1}\) of soil Pb. Defining C% as >30% also generates similar predicted and actual number of samples in the high group (high observed = 351, high predicted = 442).

<table>
<thead>
<tr>
<th>C%</th>
<th>&gt;60%</th>
<th>&gt;55%</th>
<th>&gt;50%</th>
<th>&gt;45%</th>
<th>&gt;40%</th>
<th>&gt;35%</th>
<th>&gt;30%</th>
<th>&gt;25%</th>
<th>&gt;20%</th>
<th>&gt;15%</th>
<th>&gt;10%</th>
<th>&gt;5%</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Y_I)</td>
<td>0.07</td>
<td>0.16</td>
<td>0.24</td>
<td>0.29</td>
<td>0.34</td>
<td>0.42</td>
<td><strong>0.46</strong></td>
<td>0.42</td>
<td>0.34</td>
<td>0.24</td>
<td>0.05</td>
<td>0</td>
</tr>
</tbody>
</table>

**Figure 8.9:** Receiver operating characteristic curve for Glasgow-G-BASE samples comparing the true positive rate (true positives/total positives) with the false positive rate (true negatives/total negatives) for different values of C ranging from >5% to >60%. The 30% cut-off is depicted by a red dashed line and the area under the ROC curve is added under AUC. The Youden’s index \(Y_I\) is also shown for all values of C in the table.
The observed and predicted groups were compared using a confusion matrix (Figure 8.10), and simple binary maps (Figure 8.11). The results indicate that this is a viable method for identifying locations at risk from high soil Pb across a city. The overall accuracy of the model was 74%, but more importantly the user's and producer's accuracies of the high group predictions were >50% (Figure 8.10) i.e. the majority of the samples with high soil Pb were classified as high-risk (producer's accuracy), and the majority of samples predicted to be at high-risk had high soil Pb (user's accuracy). Furthermore, the binary maps display a broadly similar pattern in locating high risk locations (Figure 8.11). This all indicates that this methodology is sufficiently accurate at identifying locations at high risk from high soil Pb contents.

<table>
<thead>
<tr>
<th>Predicted group</th>
<th>Confusion Matrix</th>
<th>User's accuracy/error</th>
<th>Producer's accuracy/error</th>
<th>Overall accuracy/error</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>High</strong> (≥30% risk of ≥200 mg kg⁻¹ Pb)</td>
<td>238 18.5%</td>
<td>217 16.9%</td>
<td>52.3% 47.7%</td>
<td>67.8% 32.2%</td>
</tr>
<tr>
<td><strong>Low</strong> (&lt;30% risk of &gt;200 mg kg⁻¹ Pb)</td>
<td>113 8.8%</td>
<td>717 55.8%</td>
<td>86.4% 13.6%</td>
<td>76.8% 23.2%</td>
</tr>
</tbody>
</table>

**Figure 8.10:** Confusion matrix comparing the predicted group i.e. samples in the Glasgow G-BASE dataset predicted to be at high risk of high soil Pb (≥30% probability of having ≥200 mg kg⁻¹ of soil Pb; mapped in Figure 8.11A) and the actual group i.e. observed G-BASE samples with high (>200 mg kg⁻¹) soil Pb (mapped in Figure 8.11B).
Figure 8.11A: Map of observed G-BASE points coloured by high (≥200 mg kg\(^{-1}\)) or low (<200 mg kg\(^{-1}\)) soil Pb.

Figure 8.11B: Map of predicted G-BASE points coloured by high (≥30% probability of having ≥200 mg kg\(^{-1}\) soil Pb) or low (<30% probability of having >200 mg kg\(^{-1}\) soil Pb) risk.
I am currently unaware of any studies attempting to identify urban locations at risk from high soil Pb via QRF, so there are no comparable studies to equate results with. The most similar studies are Schwarz et al. (2013) who used a RF model to determine if soil Pb in Baltimore (USA) would be above or below the threshold of 400 mg kg\(^{-1}\); Zhang et al. (2008) who used Local Moran’s I to find soil Pb hotspots in the Irish city of Galway; and Marchant et al. (2011) who employed robust geostatistics to quantify the input from point Pb sources in the Welsh city of Swansea. The results from this study are an improvement on Schwarz et al.’s (2013) study where the overall accuracy of their RF model was high at 72%, but they had a 50% producer’s accuracy and 25% user’s accuracy as they only classified 2/4 high samples correctly. They also had a different threshold for high soil Pb (400 mg kg\(^{-1}\)), and a smaller validation dataset (n=29) because they did not use LOOCV.

Crucially, this study expands on previous research as it determines the probability that a location will have high soil Pb. This inclusion of uncertainty is important to communicate to decision makers. Furthermore, the high soil Pb threshold and the high-risk cut-off value can both be altered if new information is found, or decision makers want different thresholds. Therefore, model adaptability is another strength of this study.

8.4.2. Neighbourhood scale predictions

This section focuses on the second research question, i.e. can a RF model accurately predict soil Pb concentrations and/or risk at the neighbourhood scale? Two neighbourhood soil Pb datasets from Paisley and Bishopbriggs were examined to answer this question.

8.4.2.1 Covariate selection

In the Paisley RF model only covariates that showed a significant relationship with soil Pb in the geostatistical analysis (Chapter 6) were included. In order of strongest relationship to weakest (based on overall Wald values) these were OM\% (log), land use, building age, building distance, and road distance (Table 6.12). According to the predictor importance plot (Figure 8.12A), all predictors have a >0.4 influence on soil Pb in the Paisley RF model. This indicates that they should all be included.

In Bishopbriggs the geostatistical analysis indicated that only three covariates had a significant relationship with soil Pb, namely OM\% (log), land use, and road distance. However, road distance had a positive relationship with soil Pb due to
high Pb samples in the middle of a park distal from roads. However, soil Pb should be elevated closer to roads due to historic leaded petrol emission, so road distance was not included in the RF model. The same is the case for building distance. However, distance to the nearest historic industry site had a negative relationship with soil Pb, which was expected. Therefore, despite historic industry distance not quite having a significant relationship with soil Pb (Wald = 3.2, $p = 0.07$, Table 6.12), it was included in the RF model. Historic industry type also had the expected relationship with soil Pb although it was not significant. Overall, four covariates were used to construct a RF model in Bishopbriggs, in order of strongest relationship to weakest (based on overall Wald values, Table 6.12) these were OM (log), land use, historic industry distance, and historic industry type. These were selected to provide enough information for the RF model to predict soil Pb, while not including covariates which have observed, unexpected relationships with soil Pb.

The predictor importance plot in Bishopbriggs suggests that some predictors have a very low but still positive (i.e. $< 0.1$), influence on soil Pb (i.e. industry distance and industry type; Figure 8.12B). However, omitting these covariates would leave only two predictors, which is too few for a RF model.
Figure 8.12: Predictor importance plots from the RF models on (A) Paisley and (B) Bishopbriggs datasets. Covariates are ordered from strongest relationship with soil Pb on the left to weakest relationship on the right (based on geostatistical analysis).
8.4.2.2. Prediction of soil Pb concentrations

The Paisley and Bishopbriggs RF models were used with QRF to predict soil Pb concentrations for a range of percentiles at a 5 by 5 m prediction grid covering each study area. First, the median log soil Pb predictions were classified into six equal intervals (Figure 8.13). Next, the 5\textsuperscript{th}, 50\textsuperscript{th}, and 95\textsuperscript{th} predicted soil Pb values were classed by the most precautionary C4SL they exceeded regardless of current land use, with darker colours for higher values (Figures 8.14 and 8.15).

The predicted soil Pb (log) maps for Paisley and Bishopbriggs are useful for discerning the spatial pattern in soil Pb predicted concentrations (Figure 8.13). Unsurprisingly, higher Pb concentrations are generally predicted along road and garden soil in Paisley (Figure 8.13A). This is because these predictors had a strong influence in the Paisley RF model (Figure 8.12A). Likewise, in Bishopbriggs higher estimated Pb concentrations correspond with public open space, and lower concentrations are predicted along road verges (Figure 8.13B). Organic matter is the most important predictor in both neighbourhood datasets, but it is difficult to estimate OM levels from a map. At each prediction grid point OM was determined by kriging known levels at sample points, which adds a degree of error. In the future, more accurate methods to determine OM should be investigated.

In the C4SL classed maps (Figures 8.14 and 8.15) the spatial pattern in predicted soil Pb concentrations is still discernible. The majority of the study areas is likely to only exceed the allotment C4SL, especially in Bishopbriggs (Figure 8.15A). In Paisley, soil around roads and garden may exceed the C4SL for garden soil (Figure 8.14A). The confidence interval maps show a similar pattern to the median prediction, but for different land uses. Some locations may exceed the C4SL for residential open space in both Paisley and Bishopbriggs (Figures 8.14C and 8.15C). Overall, the C4SL classed maps show slightly more of the variability in the dataset than in the Greater Glasgow map (Figure 8.5), maybe due to the more detailed sampling scale.
Figure 8.13: Maps of (A) Paisley and (B) Bishopbriggs (5 by 5 m prediction grid) showing median estimated soil Pb concentrations (log) predicted from a QRF algorithm, classified into six equal intervals.
Figure 8.14: Maps of Paisley grid points showing (A) median predicted soil Pb, (B) 5th percentile/lower confidence interval predictions, and (C) 95th percentile/upper lower confidence interval predictions for a 5 by 5 m prediction grid covering the study area. Determined from QRF at prediction grid points (5 by 5 m) covering the study area. Classified by the land use C4SL exceeded at each prediction grid point regardless of current land use (see Table 8.2 for values). N.B. most precautionary land use C4SL is shown, so if the point is classified as garden, the allotment C4SL is also exceeded.
Figure 8.15: Maps of Bishopbriggs grid points showing (A) median predicted soil Pb, (B) 5th percentile/lower confidence interval predictions, and (C) 95th percentile/upper confidence interval predictions for a 5 by 5 m prediction grid covering the study area. Determined from QRF at prediction grid points (5 by 5 m) covering the study area. Classified by the land use C4SL exceeded at each prediction grid point regardless of current land use (see Table 8.2 for values). N.B. most precautionary land use C4SL is shown, so if the point is classified as garden, the allotment C4SL is also exceeded.
To assess the accuracy of soil Pb predictions at the neighbourhood scale, 10-fold cross-validation was performed on the Paisley and Bishopbriggs RF models, for a range of percentiles and for all samples, and just high (Pb ≥200 mg kg\(^{-1}\)) or low samples (Pb <200 mg kg\(^{-1}\)).

For the median prediction using the whole dataset (double boxed in Table 8.6) the cross-validation statistics are good. In Paisley and Bishopbriggs the ME and RMSE are close to 0 (Paisley = -0.09 and 0.57 and Bishopbriggs = -0.17 and 0.52 respectively). The CCC is slightly higher in Paisley (0.56) than in Bishopbriggs (0.35) due to a wider prediction range, but both models show good coefficients at >0.3. Furthermore, in the median plots, the predicted samples fall either side of the 1:1 line which explains the low ME and high CCC (Figures 8.16 and 8.17).

Compared with the city scale Glasgow model, the neighbourhood models tend to perform better, although there is the same central tendency issue (Figures 8.16 and 8.17) and the neighbourhood models are less consistent. The ME and RMSE are slightly lower in the Glasgow model (ME = -0.03 and RMSE = 0.41, Table 8.5) compared with the Paisley and Bishopbriggs model, but the difference is not substantial given the smaller sample numbers in the local datasets. Furthermore, the CCC is much higher in the Paisley model (0.56) compared with Glasgow (0.33). This is reflected in the observed versus predicted plots (Figure 8.16), where the predicted values show a wider range, especially for higher values, compared with Glasgow plot (Figure 8.7). Bishopbriggs has similar CCC results to the Glasgow model, indicating not all neighbourhood scale models are more accurate than the city model.

Unsurprisingly, low soil Pb samples have the most accurate predictions at the 25\(^{th}\) percentile in the neighbourhood models, although the CCC is low (0.21 and 0.17 for Paisley and Bishopbriggs respectively) due to one or two samples still being overpredicted (Figures 8.16 and 8.17). However, low soil Pb sample accuracy has less overall influence on complete dataset compared with Glasgow, as only 58-59% of the samples have low soil Pb in Bishopbriggs or Paisley, whereas this is 73% in Glasgow.

In order to efficiently focus future sampling and remediation at locations with high soil Pb concentrations, it is more important to predict high samples. High samples are slightly overpredicted at the 95\(^{th}\) percentile, with a more accurate prediction at the 75\(^{th}\) percentile, especially in Bishopbriggs. This is reflected in the CCC, which is
0.49/0.75 at the 75th percentile but only 0.21/0.3 at the 95th (Table 8.6). However, if you wanted to be conservative then the 95th percentile predicts all samples with soil Pb higher than their actual value, whereas the 75th percentile still underpredicts soil Pb in some samples (Figures 8.16 and 8.17). Interestingly, although the Bishopbriggs model has worse cross-validation than Paisley for the whole dataset, it predicts high samples more accurately (Table 8.6). In fact, the CCC for high samples at the 75th percentile in Bishopbriggs (0.77) is the highest found in the whole analysis. This may be because there are fewer extreme high outliers in Bishopbriggs, compared with Paisley.

Importantly, the neighbourhood scale models are more accurate at predicting high Pb soil samples, compared with the city scale model (high samples CCC in Glasgow = 0.41; Table 8.5). This may be because a greater proportion of the samples in the Paisley and Bishopbriggs datasets have high soil Pb (41%, and 42% respectively), compared with the G-BASE dataset (27%). This is unsurprising as Paisley and Bishopbriggs were selected because of their high variability in soil Pb concentrations. They are also smaller datasets so will contain fewer low Pb samples from background Pb emissions. This suggests that a RF model trained on a local dataset can predict samples with high soil Pb (≥200 mg kg⁻¹) with relatively good accuracy.
Table 8.6: RF model validation statistics for (A) Paisley and (B) Bishopbriggs, including all samples, only high samples, and only low samples. The Pb values have also been predicted for a range of quantiles using QRF. Green cells = ‘best’ results. Double boxed = median prediction for all samples. The ME and RMSE values are for log transformed units.

<table>
<thead>
<tr>
<th>Soil Pb group</th>
<th>Prediction quantile</th>
<th>Mean ME</th>
<th>Mean RMSE</th>
<th>CCC</th>
<th>Pearson’s R²</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>A. PAISLEY</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>All n=112</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5th</td>
<td>-1.21</td>
<td>1.22</td>
<td>0.09</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>25th</td>
<td>-0.60</td>
<td>0.73</td>
<td>0.36</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>40th</td>
<td>-0.28</td>
<td>0.59</td>
<td>0.5</td>
<td>0.6</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>-0.09</td>
<td>0.57</td>
<td>0.56</td>
<td>0.62</td>
</tr>
<tr>
<td></td>
<td>60th</td>
<td>0.13</td>
<td>0.58</td>
<td>0.57</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>0.52</td>
<td>0.73</td>
<td>0.5</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>1.54</td>
<td>1.54</td>
<td>0.13</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>High (&gt;5.3 log Pb or &gt;200 mg kg⁻¹) n=46</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5th</td>
<td>-2.13</td>
<td>2.13</td>
<td>0.02</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>25th</td>
<td>-1.32</td>
<td>1.32</td>
<td>0.1</td>
<td>0.34</td>
</tr>
<tr>
<td></td>
<td>40th</td>
<td>-0.92</td>
<td>0.95</td>
<td>0.19</td>
<td>0.37</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>-0.69</td>
<td>0.76</td>
<td>0.27</td>
<td>0.42</td>
</tr>
<tr>
<td></td>
<td>60th</td>
<td>-0.44</td>
<td>0.6</td>
<td>0.36</td>
<td>0.44</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>-0.04</td>
<td>0.47</td>
<td>0.49</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>0.75</td>
<td>0.76</td>
<td>0.21</td>
<td>0.5</td>
</tr>
<tr>
<td></td>
<td>Low (&lt;5.3 log Pb or &lt;200 mg kg⁻¹) n=66</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5th</td>
<td>-0.58</td>
<td>0.58</td>
<td>0.11</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>25th</td>
<td>-0.11</td>
<td>0.33</td>
<td>0.21</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>40th</td>
<td>0.17</td>
<td>0.34</td>
<td>0.15</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>0.33</td>
<td>0.43</td>
<td>0.11</td>
<td>0.15</td>
</tr>
<tr>
<td></td>
<td>60th</td>
<td>0.53</td>
<td>0.56</td>
<td>0.08</td>
<td>0.14</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>0.91</td>
<td>0.91</td>
<td>0.05</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>2.09</td>
<td>2.09</td>
<td>0.01</td>
<td>0.17</td>
</tr>
<tr>
<td></td>
<td>B. BISHOPBRIGGS</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>All n=138</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5th</td>
<td>-0.97</td>
<td>0.97</td>
<td>0.04</td>
<td>0.26</td>
</tr>
<tr>
<td></td>
<td>25th</td>
<td>-0.55</td>
<td>0.66</td>
<td>0.13</td>
<td>0.54</td>
</tr>
<tr>
<td></td>
<td>40th</td>
<td>-0.33</td>
<td>0.56</td>
<td>0.22</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>-0.17</td>
<td>0.52</td>
<td>0.35</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>60th</td>
<td>0.04</td>
<td>0.51</td>
<td>0.44</td>
<td>0.58</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>0.46</td>
<td>0.60</td>
<td>0.46</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>1.28</td>
<td>1.29</td>
<td>0.13</td>
<td>0.47</td>
</tr>
<tr>
<td></td>
<td>High (&gt;5.3 log Pb or &gt;200 mg kg⁻¹) n=58</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5th</td>
<td>-1.7</td>
<td>1.7</td>
<td>0</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>25th</td>
<td>-1.24</td>
<td>1.24</td>
<td>0.04</td>
<td>0.52</td>
</tr>
<tr>
<td></td>
<td>40th</td>
<td>-0.99</td>
<td>0.99</td>
<td>0.11</td>
<td>0.59</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>-0.78</td>
<td>0.79</td>
<td>0.22</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>60th</td>
<td>-0.53</td>
<td>0.57</td>
<td>0.39</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>-0.04</td>
<td>0.32</td>
<td>0.75</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>0.69</td>
<td>0.72</td>
<td>0.3</td>
<td>0.56</td>
</tr>
<tr>
<td></td>
<td>Low (&lt;5.3 log Pb or &lt;200 mg kg⁻¹) n=80</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>5th</td>
<td>-0.45</td>
<td>0.45</td>
<td>0.07</td>
<td>0.27</td>
</tr>
<tr>
<td></td>
<td>25th</td>
<td>-0.05</td>
<td>0.24</td>
<td>0.17</td>
<td>0.23</td>
</tr>
<tr>
<td></td>
<td>40th</td>
<td>0.15</td>
<td>0.25</td>
<td>0.05</td>
<td>0.07</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>0.27</td>
<td>0.32</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>60th</td>
<td>0.45</td>
<td>0.47</td>
<td>0.02</td>
<td>-0.04</td>
</tr>
<tr>
<td></td>
<td>75th</td>
<td>0.81</td>
<td>0.81</td>
<td>0.02</td>
<td>-0.09</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>1.71</td>
<td>1.71</td>
<td>0.01</td>
<td>0.07</td>
</tr>
</tbody>
</table>
Figure 8.16: Predicted vs actual soil Pb (log) values for Paisley points using the RF and QRF for a range of percentiles (title of each plot). Points are coloured blue if their actual soil Pb is low (<200 mg kg⁻¹) and pink if actual soil Pb is ≥200 mg kg⁻¹. A 1:1 line has been added to aid interpretation. X and Y axis ranges have been fixed to match.
Figure 8.17: Predicted vs actual soil Pb (log) values for Bishopbriggs points using the RF and QRF for a range of percentiles (title of each plot). Points are coloured blue if their actual soil Pb is low (<200 mg kg⁻¹) and pink if actual soil Pb is ≥200 mg kg⁻¹. A 1:1 line has been added to aid interpretation. X and Y axis ranges have been fixed to match.
8.4.2.3. Prediction of soil Pb risk

Quantile regression forests were used to predict the probability of a sample from Paisley and Bishopbriggs exceeding 200 mg kg\(^{-1}\) of soil Pb. The resulting probabilities showed a wide range (1-97% for Paisley and 15-74% for Bishopbriggs), with a mean of 41-42%. This is much higher than the mean probability for all of Greater Glasgow (27%), but the neighbourhood datasets are smaller and were selected to include soil Pb hotspots. Additionally, Getis-Ord Gi* was used to identify probability clusters, with hotspots (>99% confidence) outlined on the probability map (Figure 8.18).

The probability maps largely agree with earlier maps (Figures 8.13 - 8.15, 5.6B and C). Specifically, there is predicted to be several soil Pb hotspots in Paisley, which often coincide with roads and houses (Figure 8.18A). In Bishopbriggs the main soil Pb hotspot is predicted to occur at Etive Crescent Park (Figure 8.18B), with smaller surrounding hotspots. This was generally what was observed on the kriged soil Pb map (Figure 5.6C).
Figure 8.18: 5 m by 5 m grid covering the study area in (A) Paisley and (B) Bishopbriggs showing the probability of a location exceeding the high soil Pb threshold of 200 mg kg⁻¹, with contour lines indicating probability hotspots (>99% confidence).
The ROC curve and Youden’s index were calculated for a range of cut-off probabilities to select the optimum cut off percentage. The ROC curve has a large under the curve area (0.95 in Paisley and 0.87 in Bishopbriggs) which is close to 1 indicating a very good model (Figure 8.19). This is higher than the AUC of the Glasgow ROC curve (0.79), suggesting the neighbourhood models will generate more accurate predictions. These AUC values are higher than those found in Obeng-Gyasi et al.’s (2021) study using a Bayesian network (AUC = 0.83) and logistic models (AUC = 0.74) to predict soil Pb in Greensboro, USA.

In Paisley, the highest Youden’s index was for a 30% cut-off value (Figure 8.19), which is the same as for Glasgow (Figure 8.9). However, the maximum Youden’s index was for a 40% cut-off value in Bishopbriggs (Figure 8.19), indicating samples with a ≥40% probability of having high soil Pb agree most closely with the observed soil Pb values. This slightly higher cut-off suggests that soil in Bishopbriggs is at a slightly higher risk of high Pb concentrations. Overall, the soil Pb risk cut-off remains similar despite different soil Pb concentrations, covariates, and range of probabilities.

<table>
<thead>
<tr>
<th>C%</th>
<th>&gt;90%</th>
<th>&gt;80%</th>
<th>&gt;70%</th>
<th>&gt;60%</th>
<th>&gt;50%</th>
<th>&gt;40%</th>
<th>&gt;30%</th>
<th>&gt;20%</th>
<th>&gt;10%</th>
<th>&gt;0%</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paisley</td>
<td>0.17</td>
<td>0.41</td>
<td>0.52</td>
<td>0.63</td>
<td>0.71</td>
<td>0.7</td>
<td><strong>0.72</strong></td>
<td>0.45</td>
<td>0.35</td>
<td>0</td>
<td><strong>0.95</strong></td>
</tr>
<tr>
<td>BB</td>
<td>0.14</td>
<td>0.24</td>
<td>0.25</td>
<td>0.39</td>
<td>0.47</td>
<td><strong>0.66</strong></td>
<td>0.45</td>
<td>0.18</td>
<td>0.01</td>
<td>0</td>
<td><strong>0.87</strong></td>
</tr>
</tbody>
</table>

**Figure 8.19:** ROC curves for Paisley and Bishopbriggs comparing the true positive rate with the false positive rate for different values of C ranging from >0% to >90%. Optimal cut-off values are shown by a dashed line. The Youden’s index is also shown for all values of C in the table, along with the AUC.
Validation of the predictions were assessed via a confusion matrix comparing observed soil Pb group i.e. high or low, with predicted soil Pb risk (Figure 8.20). It shows good overall accuracy (84% and 83%) and reasonably good user’s accuracy (73% and 76%) for the high group in Paisley and Bishopbriggs respectively (Figure 8.20). Most importantly the producer’s accuracy for high samples is very good in Paisley (98%), as only one sample with observed high soil Pb was missed (Figure 8.20). The producer’s accuracy is also good in Bishopbriggs (86%), although eight high samples were missed (Figure 8.20).

These neighbourhood scale probability predictions are more accurate than for the city scale Glasgow model (overall accuracy = 75%, high producer’s accuracy = 68%; Figure 8.10). This is likely because the neighbourhood datasets are smaller and thus quantify the relationships between soil Pb and the covariates more consistently and in more detail, to generate more accurate predictions. This is achievable over a small area but is more complex over larger areas.

The map of predicted soil Pb risk group is very similar to the map of observed soil Pb group in Paisley (Figure 8.21) and Bishopbriggs (Figure 8.22). In Paisley there are a few additionally high-risk samples in the predicted risk map (Figure 8.21B) that did not have observed high soil Pb (Figure 8.21A), but these tend to occur near to clusters of high soil Pb. This is also evident in Bishopbriggs, where hotspots on the observed soil Pb map (Figure 8.22A) have similar locations to the predicted map (Figure 8.22B) e.g. around the houses in the northwest and at the park in the northeast.

8.4.3. Practical uses

The output maps have various practical uses and end-users, with the methodology applicable to other cities, if covariate datasets are available. Firstly, the contours delineating geographical hotspots in the probability maps (Figures 8.8 and 8.18) could inform local authorities of areas most likely to be of concern for managing land contamination and remediation during city planning and development. The probability maps (Figures 8.8 and 8.18) could also be used as part of an environmental assessment desk-study before development e.g. new housing, to determine if and where field samples should be collected. The C4SLs classed maps (Figures 8.5, 8.14, and 8.15) could help decide the most appropriate future use of land for different locations in a city, e.g. do not locate new allotments on locations exceeding allotment C4SL level.
### Confusion Matrices for Paisley and Bishopbriggs

**A - Paisley**

- **High**
  - Users’ accuracy/error: 72.6% (40.2%)
  - Producer’s accuracy/error: 97.8% (2.2%)

- **Low**
  - Users’ accuracy/error: 96.0% (2.0%)
  - Producer’s accuracy/error: 74.2% (25.8%)

**B - Bishopbriggs**

- **High**
  - Users’ accuracy/error: 50 (36.2%)
  - Producer’s accuracy/error: 86.2% (13.8%)

- **Low**
  - Users’ accuracy/error: 16 (11.6%)
  - Producer’s accuracy/error: 80.0% (20.0%)

**Actual group**

- **High** (≥200 mg kg⁻¹ Pb)
- **Low** (<200 mg kg⁻¹ Pb)

**Predicted group**

- **High** (≥30% risk of ≥200 mg kg⁻¹ Pb)
- **Low** (<30% risk of >200 mg kg⁻¹ Pb)

---

**Figure 8.20:** Confusion matrices for (A) Paisley, and (B) Bishopbriggs, comparing the output class i.e. samples predicted to be at high risk of high soil Pb (≥30% probability of having ≥200 mg kg⁻¹ of Pb; mapped in Figures 8.21A and 8.22A) and the target class i.e. observed samples with high (≥200 mg kg⁻¹) soil Pb (mapped in Figures 8.21B and 8.22B).
Figure 8.21A: Map of observed Paisley samples coloured by high ($\geq 200$ mg kg$^{-1}$) or low ($< 200$ mg kg$^{-1}$) soil Pb.

Figure 8.21B: Map of predicted Paisley samples coloured by high ($\geq 30\%$ probability of having $\geq 200$ mg kg$^{-1}$ soil Pb) or low ($< 30\%$ probability of having $> 200$ mg kg$^{-1}$ soil Pb) risk.
Figure 8.22A: Map of observed Bishopbriggs samples coloured by high (≥200 mg kg⁻¹) or low (<200 mg kg⁻¹) soil Pb.

Figure 8.22B: Map of predicted Bishopbriggs samples coloured by high (≥40% probability of having ≥200 mg kg⁻¹) or low (<40% probability of having <200 mg kg⁻¹ soil Pb) risk.
8.5. Conclusions

Lead concentrations in urban soils are often difficult to predict due to over-printing from several Pb sources. Predicting areas at high-risk from high soil Pb in urban areas is crucial for future land use planning and management. However, this is extremely tricky because hotspots of soil Pb concentrations are often dependent on localised factors that do not consistently generate high Pb concentrations. Overall, it is best to err on the side of caution and over-predict soil Pb concentrations, rather than underpredict them.

This chapter has examined the accuracy and applicability of three RF models – one trained on the Glasgow city-wide dataset, and two trained on neighbourhood datasets from Paisley and Bishopbriggs. All models show sufficiently accurate predictions (Table 8.7) of both soil Pb concentrations (determined by cross-validation CCC >0.33) and soil Pb risk (assessed via confusion matrices where overall accuracy >74%). Furthermore, higher quantiles generate more accurate predictions of high soil Pb samples i.e. CCC >0.41 for high samples at the 75th/95th percentiles (Table 8.7). This is more evident in the neighbourhood scale models, which tend to be generally more accurate than the city scale model.

Additionally, different mapping methods were explored, e.g. using logarithmic values and C4SL exceedance, in order to best communicate the results to decision makers. The ‘best’ output map will ultimately depend on the end user.

This research has proven that machine learning modelling of soil Pb is possible, at different scales. Nevertheless, accuracy could be improved by exploring other covariates and machine learning methods. Urban soils are complex and heterogenous, with varied, undocumented, and inconsistent histories making soil Pb hotspots extremely difficult to model.
<table>
<thead>
<tr>
<th>Location</th>
<th>Target Variable</th>
<th>Covariates</th>
<th>Soil Pb concentrations - Cross-validation statistics</th>
<th>Soil Pb risk - Probability confusion matrix</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Samples</td>
<td>Pred. quantile</td>
</tr>
<tr>
<td>Greater Glasgow</td>
<td>log(Pb)</td>
<td>- OM (log)</td>
<td>All</td>
<td>50th</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Building age</td>
<td>high</td>
<td>95th</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- HI density (log)</td>
<td>low</td>
<td>25th</td>
</tr>
<tr>
<td>Bishopbriggs</td>
<td>log(Pb)</td>
<td>- OM (log)</td>
<td>All</td>
<td>50th</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Land use</td>
<td>high</td>
<td>75th</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- HI type</td>
<td>low</td>
<td>25th</td>
</tr>
<tr>
<td>Paisley</td>
<td>log(Pb)</td>
<td>- OM (log)</td>
<td>All</td>
<td>50th</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Land use</td>
<td>high</td>
<td>75th</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Building age</td>
<td>low</td>
<td>25th</td>
</tr>
</tbody>
</table>

Table 8.7: Summary of the RF models on the G-BASE Glasgow, Paisley and Bishopbriggs datasets, including internal 10-fold cross validation statistics for different soil Pb groups and quantiles, and confusion matrix results comparing predicted soil Pb risk with observed soil Pb group (≥ or < 200 mg kg⁻¹). * producer’s and user’s accuracies for high group samples only. HI=historic industry, CCC=concordance correlation coefficient.
References - Chapter 8

© Crown copyright and database rights (2021b) Ordnance Survey (100025252)

© Crown Copyright and Landmark Information Group Limited (2021a). All rights reserved.


Appendix 8.1 - Example random forest Matlab code

% 1. Build the RF model
% Put the covariates/predictors in a table
T2 = table (log_OM, buildingage, HI_density_log, IND_date2, LandUse, IND_type);
% Select the fixed variable (log transform due to skew)
z = log(Pb);
% Run the tree bagger algorithm
mns = 6; % max number of splits, reducing number prevents overfitting
rf2 = TreeBagger (1001, T2, z, 'Method', 'regression', 'OOBPredictorImportance','on', 'PredictorSelection', 'curvature', 'InBagFraction', 0.5, 'NumPredictorsToSample', 4, 'MaxNumSplits', mns);
% 1001 = number of trees in forest, high number and odd
% 'OOBPredictorImportance','on' means can see predictor importance
% 'InBagFraction' = fraction of samples to randomly select each time
% 'NumPredictorsToSample' = random no. of predictors to select at each split
%%%%%%%%%%%%%%%%%

% 2. Examine the RF model
% View a graph of the 900th trained tree
view(rf2.Trees{900}, 'Mode', 'graph')

% Create partial dependence plots - relationship between predicted soil Pb and a covariate, red line = partial dependence plot. Add 'Conditional', 'absolute' for ICE plot
t = tiledlayout(5,2, 'TileSpacing', 'compact');
title (t, 'PDP plots')
for i = 1 : 9
    nexttile
    plotPartialDependence(rf2, i)
    xlabel('Soil Pb (log)')
end

% Plot the out-of-bag error over the number of grown trees
oobErrorBaggedEnsemble = oobError(rf2);
plot(oobErrorBaggedEnsemble)
xlabel ('Number of grown trees');
ylabel ('Out-of-bag classification error');
min=min(oobErrorBaggedEnsemble);
txt = {'Min: ' min};
text(700, 0.38, txt);

% Extract and plot the importance of each covariate
imp = rf2.OOBPermutedPredictorDeltaError;
bplot = bar(imp);
bplot.FaceColor = '#808080';
importance = round(imp,2);
ylabel('Predictor importance estimates');
xlabel('Predictors');
text(1:length(importance),importance,num2str(importance'),'vert','bottom','horiz','center');
xticklabels({'OM (log)', 'Build age', 'Ind density (log)', 'Ind age', 'Land use', 'Ind type'});
xtickangle(45);  

%%%%%%%%%%%%%%%%%
%3. Probability a sample has high soil Pb
P_vals = 0.01:0.01:0.99; %define percentiles
Lp = length(P_vals);
% Use QRF to predict soil Pb at points for a range of percentiles
pred2 = quantilePredict(rf2,T2,'Quantile',P_vals);
% Define a threshold above which soil Pb (log) is high
threshold = log(200);
% Calculate the probability
prob = 100*sum(pred2>threshold,2)/(Lp+1);

% Classify samples into high or low based on measured soil Pb values
Pb_cat = ones(1285,1);
Pb_cat = categorical(Pb_cat);
Pb_cat(Pb>=200) = 'High';
Pb_cat(Pb<200) = 'Low';
valueset = {'High', 'Low'}; %removes empty 1 class
Pb_cat = categorical(Pb_cat,valueset);

% Reclassify columns into probability exceeds threshold
cutoff = 30; %use previously found value for C
prob2 = categorical(prob);
prob2(prob>=cutoff) = 'High';
prob2(prob<cutoff) = 'Low';
prob2 = categorical(prob2,valueset);

% Plot the confusion matrix
plotconfusion(Pb_cat,prob2)

% Plot the ROC curve
ROC=plot(False_pos,True_pos);
title('ROC - Glasgow G-BASE')
ylabel('True positive rate')
xlabel('False positive rate')
ROC.LineWidth=1;
ROC.Color = [0 0 0];
ROC.Marker = 'o';
ROC.MarkerFaceColor = [0.5,0.5,0.5];
ROC.MarkerEdgeColor = 'k';
hold on
l_x = [0.2173 0.2173];
l_y = [0 1];
line(l_x,l_y,'Color','red','LineStyle','--');
ROC_txt = {'>30%'};
text(0.22,0.8,ROC_txt,'Color','r');
AUC=round(abs(trapz(False_pos,True_pos)),2); %calculate area under curve
textAUC={'AUC= ' AUC};
text(0.8,0.1,textAUC);
hold off
% Calculate Youden’s values with probabilities to find max
youden=[prob1 (True_pos)+(1-False_pos)-1]

%Output as table with coordinates so can map in ArcGIS
output=table(X(:,1),X(:,2),prob,prob2);
writetable(output,'Prob_exceeds_Pb.csv');

% 4. 10-fold cross validation of the RF model
% WARNING takes a while to run
% Randomly allocate each measurement one of 10 folds
nfold=10; %number of folds
fold=ceil(nfold*rand(length(z),1)); %assigns each sample to one of 10 samples, ceil rounds

% Rerun model without one fold and compare predicted and actual results, repeat 10 times
mns=round(length(z)/10);
for i=1:nfold; %for each fold in turn
  f = find(fold==i); %f = only samples from fold i
  nf = find(fold~=i); %nf = samples not in fold i
  % re-run treebagger algorithm but only with samples NOT in the fold
  rf=TreeBagger(1001,T2(nf,2:end),z(nf),'Method','regression','OOB PredictorImportance','on','PredictorSelection','curvature','InBagFraction',0.5,'NumPredictorsToSample',6,'MaxNumSplits',mns);
  % Predict fold values (f) using treebagger
  rf_pred(f,1:7) = quantilePredict(rf,T2(f,1:end),'Quantile',[0.05 0.25 0.4 0.5 0.6 0.75 0.95]);
  rf_pred(nf,1:7) = quantilePredict(rf,T2(nf,1:end),'Quantile',[0.05 0.25 0.4 0.5 0.6 0.75 0.95]);
end

% Save validation results in a table
% C1=mean ME, C2=mean RMS, C3=CCC, C4=Pearson’s coefficient, C5=min, C6=max, rows - quantile prediction
valresults=ones(7,6); %create empty table to fill
for i = 1 : 7
  me_rf = rf_pred(:,i)-z; %mean error
  valresults(i,1) = mean(me_rf); %save in column 1
  rmse_rf=mean(sqrt((rf_pred(:,i)-z).^2)); %root mean square error
  valresults(i,2) = mean(rmse_rf); %save in column 2
  v1=nanvar(rf_pred(:,i)); %calc CCC
  v2=nanvar(z);
  m1=nanmean(rf_pred(:,i));
  m2=nanmean(z);
  p=corr(rf_pred(:,i),z,'rows','pairwise');
  r2_c=2*sqrt(v1)*sqrt(v2)*p/(v1+v2+(m1-m2).^2);
  valresults(i,3) = r2_c; %save in column 3
  r2=corrcoef(rf_pred(:,i),z); %calculate Pearson’s correlation
  valresults(i,4) = r2(2,1); %save in column 4
  valresults(i,5) = min(rf_pred(:,i)); %save min vale in column 5
  valresults(i,6) = max(rf_pred(:,i)); %save max vale in column 6
end
round(valresults,2) %output results rounded to 2 decimal places

%Assess accuracy of each soil Pb group i.e. high and low
T1=table(z,rf_pred); %put actual (z) and predicted values in a table
high=T1((T1.z>=5.3),:); %select high samples
low=T1((T1.z<5.3),:); %select low samples

%change group to ‘high’ to get different stats

%Run validation stats, repeat earlier analysis, but with only high/low samples
valresults_g = ones(7,6); %create empty table to fill
for i = 1 : 7 . . . etc.
round(valresults_g,2) \%output results rounded to 2 decimal places

\% Plot results coloured by soil Pb group
low_ar = table2array(low); \%convert table 2 array
low_actual = low_ar(:,1); \% select 1st column as actual Pb
low_pred = low_ar(:,[2:8]); \% select columns 2-8 as predicted Pb

high_ar = table2array(high); \%convert table 2 array to make easier
high_actual = high_ar(:,1); \% select 1st column as actual Pb
high_pred = high_ar(:,[2:7]); \% select columns 2-7 as predicted Pb

\% Plot actual vs predicted with points coloured by Pb group
t = tiledlayout(3,3,'TileSpacing','compact');
titles={['5th', '25th', '40th', '50th', '60th', '75th', '95th']};
x=[2.5 9.5];
y=[2.5 9.5];
hold 'off';
for i = 1 : 7
    nexttile
    line(x,y,'Color','k','LineStyle','--');
    hold 'on';
    s1=scatter(low_actual,low_pred(:,i),9,'c','filled','MarkerEdgeColor','k');
    s3=scatter(high_actual,high_pred(:,i),9,'m','filled','MarkerEdgeColor','k');
    title(titles(i));
xlabel('Actual soil Pb (log)')
ylabel('Predicted soil Pb (log)')
end
legend([s1,s3],['Low','High']);
set(legend,'Location','southeast');

\%5. Predict soil Pb at grid points
load Grid_GBASE; \%load grid points with matching covariates

\%Rename grid variables to match original
log_OM=log(OM);
buildingage=BA;
HI_density_log=log(HI_density);
IND_date2=IND_date3;
IND_type=IND_type2;

\% Put variables in prediction table
T2p=table(log_OM,buildingage,HI_density_log,IND_date2,LandUse,IND_type);

\% Use QRF to predict soil Pb at each grid point for a range of percentiles
pred1 = quantilePredict(rf2,T2p,'Quantile',P_vals);
prob_grid=100*sum(pred1>threshold,2)/(Lp+1);

\% Output as table with coords so can map
output = table(UT_x,UT_y,prob_grid);
writetable(output,'Predicted_prob_GBASE_grid.csv');
Chapter 9. Modelling in other cities and geographical locations

It is fundamental to ensure that a machine learning model is not only accurate where it was trained, but also at other geographical locations that it was not trained on. This checks that model has not been over-fit to the conditions in the training city i.e. Glasgow. This is important because transferability is one of the main benefits of machine learning. The model accuracy at other cities/geographical locations was tested using independent validation datasets containing soil Pb concentration and covariate values. Furthermore, the relationship between soil Pb and covariates in these datasets was investigated, in order to determine if geography controls covariate influence.

9.1. Introduction to new datasets

The urban G-BASE and Tellus projects sampled topsoil in 25 urban centres (Figure 9.1). Three BGS datasets containing soil Pb measurements were chosen to test the accuracy of the Glasgow RF model. The locations were selected based on available data i.e. some locations did not measure OM via LOI, and to give a wide geographical spread throughout the UK. Furthermore, the datasets aimed to be similar and different to Glasgow, in order to thoroughly test the RF model.

![Map of UK urban centres](image)

**Figure 9.1:** UK urban locations where topsoil was collected as part of the G-BASE and Tellus projects. From BGS, 2022.
Firstly, Belfast (Tellus Project, Young and Donald, 2013; Figure 9.2A) was selected as it is a large city that was home to extensive heavy historic industry i.e. shipbuilding, so has a similar history to Glasgow. Secondly, Leicester (G-BASE, Flight and Scheib., 2011; Figure 9.2B) was chosen as it is an old Roman city but had less historic heavy industry compared with Glasgow. Lastly the survey of the River Tamar catchment in the south-west of England (G-BASE, Rawlins et al., 2003; Figure 9.2C) was examined as it is a rural location so the controls on soil Pb will likely be different to those in Glasgow; however, there is extensive historic metal mining in the region. Furthermore, soil Pb in the surrounding region has already been predicted using QRF by Kirkwood et al. (2016). Although the Kirkwood et al. (2016) study does not use the same dataset, the locations are similar so model comparison will be interesting.

Despite being collected under different projects e.g. Tellus and G-BASE, the same methodology (Johnston et al., 2005) was used to collect and analyse all samples, so the results are comparable. However, the sampling density varied slightly, with a lower density of samples collected from the rural Tamar catchment (Table 9.1).

<table>
<thead>
<tr>
<th>Location</th>
<th>Total samples</th>
<th>Total area (km²)</th>
<th>Samples per km²</th>
<th>Mean Pb (mg kg⁻¹)</th>
<th>Survey date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glasgow</td>
<td>1285</td>
<td>289</td>
<td>4.45</td>
<td>199</td>
<td>2001/2</td>
</tr>
<tr>
<td>Belfast</td>
<td>866</td>
<td>235</td>
<td>3.69</td>
<td>113</td>
<td>2006</td>
</tr>
<tr>
<td>Leicester</td>
<td>656</td>
<td>170</td>
<td>3.86</td>
<td>114</td>
<td>1999</td>
</tr>
<tr>
<td>Tamar</td>
<td>468</td>
<td>956</td>
<td>0.49</td>
<td>49</td>
<td>2012</td>
</tr>
</tbody>
</table>

External testing of a machine learning model on an area outside where the model was trained is tricky and thus is absent from all other studies predicting soil Pb values (Table 8.1). Soil Pb prediction is extremely difficult using a model from a different region because geographical location can often influence the controls on soil Pb due to differences in city development and features. Furthermore, the level of historic diffuse Pb pollution varies between cities, meaning the general background concentrations of Pb varies between urban soils. Higher mean soil Pb concentrations were observed in Glasgow (199 mg kg⁻¹) compared with Belfast (113 mg kg⁻¹), Leicester (114 mg kg⁻¹), and especially Tamar (49 mg kg⁻¹). Consequently, the Glasgow RF model will expect background soil Pb concentrations to be higher in the other cities and geographical locations.
Figure 9.2A: Tellus sample locations in Belfast.
Figure 9.2B: G-BASE sample locations in Leicester.
Figure 9.2C: G-BASE sample locations in the Tamar catchment.
9.2. **Chapter aims**

This chapter aims to answer two research questions:

1. Does geographical location impact RF accuracy, at the city and neighbourhood scale?
2. Does geographical location impact the influence of covariates on soil Pb concentrations at the city scale?

9.3. **Methodology**

Firstly, each study area was defined and samples falling within that area were tidied up, to correspond with the Glasgow dataset. Subsequently, covariate values (soil OM, land use, building age, and historic industry density, age, and type) were determined at sample points (Figure 9.2), but also at each point on a 200 m by 200 m prediction grid covering the study area. This was achieved using same methodology as previous (summarised in Table 8.3), with covariate values carefully formatted to match the Glasgow dataset.

A similar methodology as outlined in Section 8.3. was used to determine the accuracy of a RF model in a region it was not trained in (see Appendix 9.1 for example code). Specifically, the Glasgow QRF model was used to predict soil Pb (log) concentrations for each percentile at every sample point in Belfast, Leicester, and Tamar. The predicted and observed values were then compared, for all points and just high (≥ 200 mg kg\(^{-1}\)) or low (< 200 mg kg\(^{-1}\)) Pb samples over a range of percentiles. Additionally, the percentile which exceeded the threshold value (200 mg kg\(^{-1}\)) was extracted and multiplied by 100 to give a risk probability percentage. These probabilities were mapped, and hotspots were highlighted using Getis-Ord Gi * cluster analysis. Furthermore, the predicted risk (high or low) was compared with observed values in a confusion matrix. This process was repeated in Bishopbriggs and Paisley, where the alternative neighbourhood dataset was used as an external validation dataset.

In order to determine the impact of geographical location on covariate influence, two new RF models were trained on soil Pb concentrations for samples in Belfast and Leicester. The relative importance of each covariate was plotted, along with predictor dependence plots. Furthermore, summary statistics for continuous variables, and the number of samples in each class for categorical variables, were calculated in every study dataset.
9.4. Results

9.4.1. Prediction of soil Pb concentrations – city scale

The Glasgow city QRF model was used to predict soil Pb concentrations at GBASE/Tellus sample locations in Belfast, Leicester, and the Tamar catchment. These predicted values were compared with actual values using validation statistics (Table 9.2) and scatterplots (Figure 9.3).

<table>
<thead>
<tr>
<th>Soil Pb group</th>
<th>Prediction Percentile</th>
<th>ME</th>
<th>RMSE</th>
<th>CCC</th>
<th>Pearson’s R²</th>
<th>Soil Pb (log) range</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>BELFAST</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All (n=866)</td>
<td>25th</td>
<td>0.16</td>
<td>0.6</td>
<td>0.16</td>
<td>0.26</td>
<td>3.76 - 5.27</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>0.5</td>
<td>0.74</td>
<td>0.13</td>
<td>0.27</td>
<td>4.25 - 5.62</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>1.63</td>
<td>1.66</td>
<td>0.05</td>
<td>0.26</td>
<td>5.27 - 7.18</td>
</tr>
<tr>
<td>Low (n=768)</td>
<td>25th</td>
<td>0.33</td>
<td>0.52</td>
<td>0.14</td>
<td>0.24</td>
<td>3.76 - 5.27</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>0.67</td>
<td>0.73</td>
<td>0.08</td>
<td>0.23</td>
<td>4.25 - 5.62</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>1.79</td>
<td>1.79</td>
<td>0.02</td>
<td>0.21</td>
<td>5.27 - 7.18</td>
</tr>
<tr>
<td>High (n=98)</td>
<td>25th</td>
<td>-1.21</td>
<td>1.21</td>
<td>0</td>
<td>0.02</td>
<td>3.79 - 5.22</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>-0.86</td>
<td>0.86</td>
<td>0.01</td>
<td>0.05</td>
<td>4.26 - 5.58</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>0.33</td>
<td>0.59</td>
<td>0.08</td>
<td>0.1</td>
<td>5.35 - 6.95</td>
</tr>
<tr>
<td>B</td>
<td>LEICESTER</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All (n=656)</td>
<td>25th</td>
<td>0.14</td>
<td>0.58</td>
<td>0.21</td>
<td>0.42</td>
<td>4.07 - 5.22</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>0.49</td>
<td>0.72</td>
<td>0.18</td>
<td>0.46</td>
<td>4.45 - 5.58</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>1.63</td>
<td>1.65</td>
<td>0.06</td>
<td>0.4</td>
<td>5.41 - 7.05</td>
</tr>
<tr>
<td>Low (n=579)</td>
<td>25th</td>
<td>0.32</td>
<td>0.5</td>
<td>0.17</td>
<td>0.35</td>
<td>4.07 - 5.22</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>0.66</td>
<td>0.71</td>
<td>0.1</td>
<td>0.38</td>
<td>4.45 - 5.58</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>1.8</td>
<td>1.8</td>
<td>0.03</td>
<td>0.35</td>
<td>5.41 - 7.05</td>
</tr>
<tr>
<td>High (n=77)</td>
<td>25th</td>
<td>-1.19</td>
<td>1.19</td>
<td>0.03</td>
<td>0.16</td>
<td>4.19 - 5.19</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>-0.82</td>
<td>0.82</td>
<td>0.05</td>
<td>0.18</td>
<td>4.56 - 5.55</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>0.35</td>
<td>0.54</td>
<td>0.08</td>
<td>0.11</td>
<td>5.56 - 6.98</td>
</tr>
<tr>
<td>C</td>
<td>TAMAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>All (n=468)</td>
<td>25th</td>
<td>0.84</td>
<td>0.9</td>
<td>0.05</td>
<td>0.28</td>
<td>4.17 - 4.87</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>1.17</td>
<td>1.2</td>
<td>0.03</td>
<td>0.29</td>
<td>4.54 - 5.27</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>2.33</td>
<td>2.33</td>
<td>0.01</td>
<td>0.31</td>
<td>5.53 - 6.87</td>
</tr>
<tr>
<td>Low (n=459)</td>
<td>25th</td>
<td>0.87</td>
<td>0.9</td>
<td>0.04</td>
<td>0.28</td>
<td>4.17 - 4.87</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>1.2</td>
<td>1.21</td>
<td>0.03</td>
<td>0.3</td>
<td>4.54 - 5.27</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>2.37</td>
<td>2.37</td>
<td>0.01</td>
<td>0.32</td>
<td>5.53 - 6.87</td>
</tr>
<tr>
<td>High (n=9)</td>
<td>25th</td>
<td>-1.09</td>
<td>1.09</td>
<td>0.02</td>
<td>0.31</td>
<td>4.42 - 4.8</td>
</tr>
<tr>
<td></td>
<td>50th</td>
<td>-0.76</td>
<td>0.76</td>
<td>0.03</td>
<td>0.22</td>
<td>4.76 - 5.17</td>
</tr>
<tr>
<td></td>
<td>95th</td>
<td>0.46</td>
<td>0.46</td>
<td>0.06</td>
<td>0.18</td>
<td>5.89 - 6.47</td>
</tr>
</tbody>
</table>
Figure 9.3: Predicted versus actual soil Pb (log) values for (A) Belfast and (B) Leicester G-BASE/Tellus points using the Glasgow RF and QRF model, for a range of percentiles (title of each plot). Points are coloured blue if their actual soil Pb is low (<200 mg kg⁻¹) and pink if actual soil Pb is high (≥200 mg kg⁻¹). A 1:1 line has been added to aid interpretation. X and Y axis ranges have been fixed to match.
9.4.2. Prediction of soil Pb risk – city scale

Quantile regression forests were combined with probability analysis to identify sample points in Belfast, Leicester, and Tamar at high risk from soil Pb. Locations with a $>30\%$ probability of having high soil Pb (based on previously defined cut-off percentage) were classed as ‘high-risk’ and compared with observed high-Pb samples (i.e. samples with $\geq 200$ mg kg$^{-1}$ of soil Pb) using a confusion matrix (Table 9.3) and maps (Figures 9.4 - 9.6).

<table>
<thead>
<tr>
<th>Location</th>
<th>Overall accuracy</th>
<th>High producer’s accuracy</th>
<th>High user’s accuracy</th>
<th>Number of high predicted</th>
<th>Number of high observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glasgow</td>
<td>74%</td>
<td>68%</td>
<td>53%</td>
<td>454</td>
<td>351</td>
</tr>
<tr>
<td>Belfast</td>
<td>71%</td>
<td>53%</td>
<td>20%</td>
<td>255</td>
<td>108</td>
</tr>
<tr>
<td>Leicester</td>
<td>80%</td>
<td>47%</td>
<td>29%</td>
<td>125</td>
<td>77</td>
</tr>
<tr>
<td>Tamar</td>
<td>69%</td>
<td>67%</td>
<td>4%</td>
<td>147</td>
<td>9</td>
</tr>
<tr>
<td>Average</td>
<td>74%</td>
<td>59%</td>
<td>27%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In Belfast the observed (Figure 9.4A) and predicted (Figure 9.4B) high-risk G-BASE samples match in places. However, there is an observed soil Pb hotspot northeast of the city centre (Figure 9.4A), which is smaller on the predicted risk map (Figure 9.4B). Furthermore, predicted high-risk clusters around old houses/historic industry (Figure 9.4B) are often more extensive than observed values (Figure 9.4A).

In Leicester the predicted (Figure 9.5B) and observed (Figure 9.5A) high-risk locations are more similar. The high-risk sites in both maps are unsurprisingly clustered around the city centre, which has older housing and a higher density of historic industry. Other high-risk sites tend to be located close to clusters of historic heavy industrial sites or old outlying towns, with the predicted risk map (Figure 9.5B) often forming tighter clusters around locations with similar covariates.

There is even less agreement between observed (Figure 9.6A) and predicted (Figure 9.6B) high-risk soil Pb sites in the Tamar dataset. The few sites with high observed soil Pb are all clustered around old mines in the south of the study area (Figure 9.6A). Subsequently, there are more scattered locations predicted to be at high-risk from soil Pb (Figure 9.6B), that are generally located close to historic industrial sites.
**Figure 9.4A:** Map of *observed* Belfast G-BASE points coloured by high (≥200 mg kg$^{-1}$) or low (<200 mg kg$^{-1}$) soil Pb.

**Figure 9.4B:** Map of *predicted* Belfast G-BASE points coloured by high (≥30% probability of having ≥200 mg kg$^{-1}$ soil Pb) or low (<30% probability of having ≥200 mg kg$^{-1}$ soil Pb) risk.
**Figure 9.5A:** Map of observed Leicester G-BASE points coloured by high (≥200 mg kg\(^{-1}\)) or low (<200 mg kg\(^{-1}\)) soil Pb.

**Figure 9.5B:** Map of predicted Leicester G-BASE points coloured by high (≥30\% probability of having ≥200 mg kg\(^{-1}\) soil Pb) or low (<30\% probability of having ≥200 mg kg\(^{-1}\) soil Pb) risk.
Figure 9.6A: Map of observed Tamar G-BASE points coloured by high (≥200 mg kg\(^{-1}\)) or low (<200 mg kg\(^{-1}\)) soil Pb.

Figure 9.6B: Map of predicted Tamar G-BASE points coloured by high (≥30% probability of having ≥200 mg kg\(^{-1}\) soil Pb) or low (<30% probability of having ≥200 mg kg\(^{-1}\) soil Pb) risk.
Soil Pb risk could also be mapped for an unmeasured city. This is achieved by creating a prediction grid for the study area, which was spatially joined with covariate layers. Subsequently, the Glasgow QRF model was used to determine the potential risk of each location having high soil Pb concentrations. Hotpot analysis i.e. Getis-Ord Gi* then located potential high-risk Pb hotspots, where >95% confidence hotspots were contoured to highlight the most at-risk locations. This was done for Leicester to illustrate the potential for identifying high-risk soil Pb locations throughout an entire city (Figure 9.7).

The predicted probability map in Leicester indicates that the areas most at risk from high soil Pb concentrations occur in the city centre and a small area in the southeast of the city (Figure 9.7). This largely agrees with locations where soil Pb has been measured at ≥200 mg kg⁻¹ (Figure 9.5A). As before, these hotspots tend to occur around areas with older (pre-1920) housing and a higher density of historic industry.

One issue is that the extent of some covariate layers may not cover the entire study area, for example the OS Greenspace layer which defines land use had a slightly reduced extent in Leicester. To solve this issue, only grid points with details for all six covariates were included. Organic matter also poses a problem as there is no digital dataset available for large areas. In Leicester OM percentages at grid points were determined by kriging known sample points to interpolate OM at unknown locations. This does add a degree of error and requires samples be collected. Alternatively, some spectral bands have been shown to correlate with soil OM; therefore, remote sensing data could be used to infer soil OM levels over a wide area (Ayoubi et al., 2011). Wang et al. (2018) achieved this using hyperspectral and Landsat data to predict soil OM in bare soil in northwest China. Lastly, problematic covariates could be removed from the RF model to make it easier to run on other cities; however, this would reduce the accuracy of model predictions.
Figure 9.7: Map showing the probability of high soil Pb (≥200 mg kg⁻¹) for a grid covering Leicester. Some grid points are missing due to missing covariate values at these locations. Contour lines indicate probability hotspots (>95% confidence).
9.4.3. Prediction of soil Pb concentrations and risk - neighbourhood scale

The Paisley QRF model was used to predict samples’ soil Pb values in the Bishopbriggs dataset (Table 9.4A), and those at high-risk of high soil Pb concentrations (Figure 9.8A). This was also done in reverse i.e. the Bishopbriggs QRF model was used to predict soil Pb concentrations (Table 9.4B) and probability of high Pb (Figure 9.8B) in the Paisley dataset. These predictions were compared with the observed soil Pb values and groups.

<table>
<thead>
<tr>
<th>Quantile</th>
<th>Mean ME</th>
<th>Mean RMSE</th>
<th>CCC</th>
<th>Pearson’s R²</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>A - Bishopbriggs values predicted from the Paisley QRF model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
<td>-1.16</td>
<td>1.16</td>
<td>0.02</td>
<td>0.15</td>
</tr>
<tr>
<td>50&lt;sup&gt;th&lt;/sup&gt;</td>
<td>-0.87</td>
<td>0.91</td>
<td>0.04</td>
<td>0.27</td>
</tr>
<tr>
<td>75&lt;sup&gt;th&lt;/sup&gt;</td>
<td>-0.38</td>
<td>0.77</td>
<td>-0.08</td>
<td>-0.12</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.51</td>
<td>0.97</td>
<td>-0.09</td>
<td>-0.13</td>
</tr>
<tr>
<td><strong>B - Paisley values predicted from the Bishopbriggs QRF model</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.95</td>
<td>1.11</td>
<td>0.21</td>
<td>0.41</td>
</tr>
<tr>
<td>50&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1.44</td>
<td>1.48</td>
<td>0.12</td>
<td>0.39</td>
</tr>
<tr>
<td>75&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1.83</td>
<td>1.83</td>
<td>0.08</td>
<td>0.39</td>
</tr>
<tr>
<td>95&lt;sup&gt;th&lt;/sup&gt;</td>
<td>2.34</td>
<td>2.34</td>
<td>0.05</td>
<td>0.37</td>
</tr>
</tbody>
</table>

*Table 9.4: External validation results from predicting soil Pb values for a range of quantiles. Green cells – ‘best’ prediction. Double boxed – median prediction. The ME and RMSE are for log transformed units.*
**Figure 9.8**: External validation confusion matrices for **A** - Bishopbriggs samples predicted using the Paisley RF model, and **B** - Paisley samples predicted using the Bishopbriggs model, comparing the output class i.e. samples predicted to be at high risk of high soil Pb (≥30% probability of having ≥200 mg kg⁻¹ of Pb) and the target class i.e. observed samples with high (≥200 mg kg⁻¹) soil Pb.
9.4.4. Covariate influence – city scale

The relative importance of each covariate is shown in the predictor importance plots (Figure 9.9). With predictor dependence plots (PDP; Figures 9.10 and 9.11) giving more detail about the relationship between soil Pb and each covariate. Furthermore, summary statistics for continuous variables, and the number of samples in each class for categorical variables, were calculated in every G-BASE sample point (Table 9.5).

**Figure 9.9:** Predictor importance plots (PIP) from RF models run on each G-BASE dataset. Glasgow is a repeat of Figure 8.3.
<table>
<thead>
<tr>
<th>Location</th>
<th>GLASGOW</th>
<th>BELFAST</th>
<th>LEICESTER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total samples</td>
<td>1285</td>
<td>866</td>
<td>656</td>
</tr>
</tbody>
</table>

**A - Soil Pb (mg kg⁻¹)**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (median)</td>
<td>199 (135)</td>
<td>113 (69)</td>
<td>114 (65)</td>
</tr>
<tr>
<td>Range</td>
<td>17 - 9937</td>
<td>11 - 3467</td>
<td>16 - 2636</td>
</tr>
</tbody>
</table>

**B - OM (%)**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (median)</td>
<td>9.55 (9.05)</td>
<td>9.83 (9.03)</td>
<td>13.04 (12.03)</td>
</tr>
<tr>
<td>Range</td>
<td>2.42 - 36.25</td>
<td>1.49 - 44.08</td>
<td>5.08 - 61.03</td>
</tr>
</tbody>
</table>

**C - Building age**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-1920</td>
<td>46%</td>
<td>15%</td>
</tr>
<tr>
<td>1920-70</td>
<td>23%</td>
<td>14%</td>
</tr>
<tr>
<td>post-1970</td>
<td>21%</td>
<td>28%</td>
</tr>
</tbody>
</table>

**D - Historic industry density (industries per km²)**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean (median)</td>
<td>2.43 (1.37)</td>
<td>1.24 (0.86)</td>
<td>0.59 (0.42)</td>
</tr>
<tr>
<td>Range</td>
<td>0.09 - 18.63</td>
<td>0.01 - 8.93</td>
<td>0 - 2.79</td>
</tr>
</tbody>
</table>

**E - Historic industry age**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N/A</td>
<td>869</td>
<td>726</td>
<td>609</td>
</tr>
<tr>
<td>Pre-1920</td>
<td>315</td>
<td>57</td>
<td>10</td>
</tr>
<tr>
<td>1920-70</td>
<td>76</td>
<td>54</td>
<td>22</td>
</tr>
<tr>
<td>Post-1970</td>
<td>25</td>
<td>29</td>
<td>15</td>
</tr>
</tbody>
</table>

**F - Land use**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Natural</td>
<td>Commercia l</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Road V.</td>
<td>Res. OS</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Garden</td>
<td>Open space</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>33%</td>
<td>4%</td>
<td>13%</td>
</tr>
<tr>
<td></td>
<td>13%</td>
<td>25%</td>
<td>20%</td>
</tr>
<tr>
<td></td>
<td>13%</td>
<td>18%</td>
<td>16%</td>
</tr>
</tbody>
</table>

**G - Historic industry type**

<p>| | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Light H.I.</td>
<td>234</td>
<td>104</td>
<td>37</td>
</tr>
<tr>
<td>Heavy H.I.</td>
<td>128</td>
<td>36</td>
<td>10</td>
</tr>
<tr>
<td>Non-H.I.</td>
<td>923</td>
<td>726</td>
<td>609</td>
</tr>
</tbody>
</table>
Figure 9.10: Predictor dependence plots showing the relationship between soil Pb (log) and (A) OM (log), (B) historic industry density (log), (C) historic industry type, and (D) land use in the RF models created for each of the three cities.
Figure 9.11: Digitised building age polygons for Glasgow, Belfast, and Leicester, with pre-1920s buildings in pink, 1920-1970 buildings in blue, and post-1970 buildings in purple. Also shown is the G-BASE study area coverage as a black line. The building age PDP from a RF on each city is also included below to give an idea of average soil Pb levels for each building age group, note different scales on the y axis.
9.5. Discussion

9.5.1. The importance of geographical location on RF accuracy

Validation of predicted soil Pb concentrations in Belfast, Leicester, and the Tamar catchment generated surprisingly good results. Specifically, the median predicted soil Pb concentrations has a relatively good agreement if only the Pearson’s correlation coefficient is examined, as the $R^2$ value is close to or $>0.3$ (for all samples, median prediction, $R^2$ value = 0.27 in Belfast, 0.46 in Leicester, and 0.29 in the Tamar dataset; Table 9.2). These values are slightly lower than the cross-validation Pearson’s $R^2$ for median predictions of all samples in the Glasgow GBASE dataset (0.51, Table 8.5), but this was to be expected as the model was built using the conditions in Glasgow. However, the Pearson’s correlation coefficient is not such a robust measure of correlation as it does not assess how far the predictions are from the 1:1 line, instead the CCC will be used, which also shows closer agreement with the ME and RMSE (Table 9.2).

The CCC is 0.1-0.2 for the city datasets (i.e. Belfast and Leicester), but much lower at $<0.1$ for Tamar, which is a rural area (Table 9.2). These CCC are much lower than the Pearson’s correlation coefficients. This suggests that the Glasgow QRF model has some capability to predict where highs and lows occur in other cities, but the magnitude of the predictions is biased due to differences between cities that cannot easily be included in covariates. Furthermore, for the whole dataset the highest CCC and lowest ME and RMSE values are for the 25th percentile predictions in both Leicester and Belfast (Table 9.2; Figure 9.3). This is likely because Glasgow has on average higher soil Pb concentrations compared with Belfast and Leicester, in these cities $>88\%$ of samples have low soil Pb values ($<200$ mg kg$^{-1}$), whereas only 73% of samples have low soil Pb concentrations in Glasgow. Therefore, the majority of the samples are more accurately predicted using a lower percentile, meaning the overall validation statistics are better for this percentile (Figure 9.3).

Conversely, samples with high ($\geq 200$ mg kg$^{-1}$) soil Pb concentrations are more accurately predicted in Belfast and Leicester using the 95th percentiles (Table 9.2; Figure 9.3), as was seen in cross-validation of the Glasgow dataset (Table 8.5). Although the CCC is only 0.1-0.11 for high samples predicted using the 95th percentile in Belfast and Leicester, due to extremely high samples remaining underpredicted at this percentile (Table 9.2; Figure 9.3). These samples fall closer
to the 1:1 line at the 99th percentile. This could be due to the central tendency issue of RFs, or that different covariates cause extreme outliers, or simply that point Pb pollution is difficult to model. However, a CCC that is positive and >0.1 indicates that the Glasgow QRF model does a reasonably good job of predicting high soil Pb concentrations in other cities, which is very difficult to achieve.

When examining the confusion matrices which assess the accuracy of soil Pb high-risk prediction (Table 9.3) the most important results are the producer’s and user’s accuracy of the high-risk groups. This is because it is more important to correctly identify potential soil Pb hot-spots rather than low-spots. Surprisingly, the Tamar dataset performs the ‘best’ in terms of producer’s accuracy (67%); however, there were only nine actual high samples and 141 false positives causing a very low user’s accuracy (4%; Table 9.3). Belfast and Leicester have a better balance with around 50% producer’s accuracy for the high group (Table 9.3), indicating half the samples with high soil Pb values were predicted to be high-risk. This is only 15-21% lower than the Glasgow cross-validation results (Table 9.3).

The users’ accuracies were slightly lower in Belfast and Leicester (20% and 29% respectively; Table 9.3), but these values are still approximately 20 - 30% lower than the Glasgow results. The low user’s accuracy is predominantly due to over-prediction of high-risk areas i.e. a large number of false positives, which may be due to soil Pb concentrations in Glasgow being generally higher than the other cities. This suggests that the ‘risk’ level may need to be lowered in some cities. However, this indicates that the model is being conservative, which is required when dealing with potential health issues from soil Pb.

Crucially, this study expands on previous research as it tests the model on other cities and geographical regions and determines the probability that a location will have high soil Pb. This inclusion of uncertainty is important to communicate to decision makers. Furthermore, the high soil Pb threshold and the high-risk cut-off value can both be altered if new information is found, or decision makers want different thresholds.

At the neighbourhood scale geographical location has an inconsistent impact on RF accuracy. For Bishopbriggs predictions using the Paisley QRF model, the resulting external validation statistics are not good; the CCC for the median prediction is close to 0 at 0.04 (although the Pearson’s $R^2$ is better at 0.27; Table 9.4A), and the overall accuracy of the risk predictions is only 53% (Figure 9.8A).
More concerning is that the high group’s producer’s accuracy is 12% i.e. only seven samples are predicted as high-risk, and 51 samples are missed (Figure 9.8A).

These poor external validation statistics of the Bishopbriggs samples using the Paisley QRF model may be due to by two reasons. Firstly, soil Pb concentrations tend to be slightly higher in Bishopbriggs compared with Paisley, median soil Pb concentrations are 170 mg kg\(^{-1}\) and 151 mg kg\(^{-1}\) respectively. Therefore, the Paisley model has been built to handle slightly lower soil Pb values, so it will be less likely to predict high-risk samples. Secondly, previous analysis has indicated that the relationship between soil Pb and the covariates varies between Paisley and Bishopbriggs (see Chapter 4). Consequently, the Paisley model has been overfit to the conditions in Paisley and cannot generalise to accurately predict soil Pb levels in other Glasgow neighbourhoods.

Conversely, for Paisley predictions using the Bishopbriggs model, the resulting external validation statistics are reasonable. The median CCC is 0.12 (Table 9.4B), and although the overall accuracy is only 44%, the high group producer’s accuracy is 100% and the user’s accuracy is 42% in the confusion matrix (Figure 9.8B). Furthermore, the most accurate soil Pb predicted values occur for the lowest percentile (25\(^{th}\)), where the CCC is good at 0.21 (Table 9.4B).

This is a significant improvement, compared with the Paisley model predictions for Bishopbriggs samples, and is likely due to higher soil Pb values in Bishopbriggs meaning the Bishopbriggs model will somewhat overpredict the Pb concentration or group, leading to a high producer’s accuracy but low user’s accuracy. This is also supported by the fact that the ‘best’ prediction is for the lowest percentile (25\(^{th}\)), indicating that soil Pb values in Paisley are overpredicted at the median, when using the Bishopbriggs model.

### 9.5.2. The importance of geographical location on covariate influence

The second aim of this chapter was to determine if geographical location influences the relationship between covariates and soil Pb in different urban environments. This was first explored by comparing predictor importance and dependence plots for RF models run on the datasets from Glasgow, Belfast, and Leicester.

One of the main differences between covariate importance in the G-BASE urban datasets (Figure 9.9) is that OM is considerably less important in Belfast (ranked 2\(^{nd}\))
and Leicester (ranked 4th), compared with Glasgow (ranked 1st; Figure 9.9). This is not linked to the range of OM values as Glasgow samples have a smaller range (2-36%) and a similar median (9%) compared with Belfast (range = 1-44%; median = 9%) and Leicester (range = 5-61%; median = 12%; Table 9.5B). Alternatively, OM may have more influence in Glasgow because OM is a sink not a source of Pb, thus is dependent on the amount of Pb deposited to the soil. Glasgow’s heavy industrial past means diffuse soil Pb pollution may have been higher, compared with Leicester and Belfast. This is proven as soils in Glasgow have higher median (135 mg kg\(^{-1}\) vs 65/69 mg kg\(^{-1}\)) and maximum (9,937 mg kg\(^{-1}\) vs 2,636/3,467 mg kg\(^{-1}\)) Pb concentrations than in Leicester and Belfast (Table 9.5A). This can also be seen on the PDP plots, where the range of soil Pb in Glasgow is much higher than in Leicester and Belfast, over the same range in OM, despite the general pattern matching (Figure 9.10A).

Building age is the most important covariate in both Belfast and Leicester and is an important covariate in the Glasgow model (Figure 9.9). Building age is a crucial covariate in all models because the cities have similar developmental histories. Firstly, all the cities have ancient routes, i.e. there is evidence that the site of Glasgow was inhabited pre-Roman era (Fry, 2017); Leicester is one of the oldest cities in England also dating to before the Roman era (Hoskins, 1951); and Belfast has Bronze age history (Beckett, 1983). Secondly, all the cities experienced substantial growth during the industrial revolution (mid-19th C), when rapid urbanisation caused new houses and factories to spring up around the city’s medieval nucleus and outlying towns (Figure 9.11). Soils around these old buildings have the highest Pb levels in all the cities (Figure 9.11) due to the use of Pb-based paint, dumping of coal ash, a long history of leaded petrol emissions, and diffuse industrial Pb pollution. Interestingly, the actual mean level of soils around pre-1920s buildings varies; with the highest value in Glasgow (~5.1 log Pb) and the lowest in Belfast (~4.6 log Pb; Figure 9.11). This shows how local factors e.g. background Pb emissions, can influence the range of Pb concentrations in different building age categories.

Later (1920-70) buildings have lower Pb levels in the surrounding soil compared with pre-1920s buildings in all cities (Figure 9.11). Although Belfast has the smallest difference in soil Pb values from around pre-1920 and 1920-70 buildings, perhaps due to later industrial development. Generally lower soil Pb concentrations around
buildings constructed between 1920 and 1970 is likely due to new environmental
laws such as the Town and Country Planning Act of 1947, which required that
planning permission had to be obtained before new building (Christman, 2013).
Constraints on building meant new houses tended to be further from industrial sites
and were less crowded with more greenspaces (Christman, 2013). Increased
planning may have resulted in more 1920-1970s buildings on less polluted Pb soils,
but leaded petrol and paints were widely used during the 20th century, and thus will
have enhanced Pb levels in some of these soils.

A small proportion of the soil samples are <5 m from post-1970 developments (11-
14%; Table 9.5C). Stricter environmental laws at the end of the 20th century e.g. the
Control of Pollution Act and the Environmental Protection Act both of 1974, caused
a slowdown in pollution. For example, the aforementioned laws, the ban on leaded
petrol in 1999 (Oudijk, 2010), the ban on Pb addition to paint in 1992 (Turner and
Sogo, 2012), and the popularisation of central heating from the 1970s causing less
coal ash dumping (Thornton, 1991). Therefore, these samples have the lowest
median soil Pb levels in each location (Figure 9.11), due to fewer Pb sources. They
also tend to be located around satellite towns that were built on greenfield sites so
were less likely to be subject to historic pollution (Figure 9.11).

Historic industry density is also an important and consistent control on soil Pb in all
three cities; it is ranked 2nd in Glasgow, Belfast, and Leicester (Figure 9.9) and shows
a similar relationship with soil Pb values in each city (Figure 9.10B). However, the
other historic industry covariates i.e. type, age, and distance, generally have little
control on soil Pb levels in all the models (Figure 9.9). This suggests that the density
of historic industry is more important than its type, age, or distance.

The type and amount of historic industry varies between the cities and may
influence its relationship with soil Pb. For example, from the 19th C Belfast was
renowned for its shipbuilding and associated industries i.e. engineering, and rope
works (Lowry, 1913; Lambert, 2021a). However, it was also home to other industries
such as printing/paper making (from 1749), distilleries and tobacco making (late
19th C), and linen weaving and dyeing (Lambert, 2021a). Glasgow had a similar
industrial past to Belfast, with textiles – mainly cotton weaving and bleaching –
dominating in the 18th C (Allison et al., 2004). In the 19th and 20th centuries metal
working and heavy engineering took forefront, which included shipbuilding
(Allison et al., 2004). This occurred on a larger scale than in Belfast, hence the
higher industrial density (average = 2.43 vs 1.24 industries per km$^2$), and higher soil Pb levels (average = 199 vs 113 mg kg$^{-1}$) in Glasgow compared with Belfast (Table 9.5D and A). Leicester’s industrial history was different from that of Glasgow and Belfast. The main industries in Leicester were hosiery and footwear manufacturing, with 2,741 boot and shoemakers in Leicester in 1861 (Lambert, 2021b). Engineering and ironworks were also present in Leicester in the 19th C, but on a much smaller scale than in Glasgow (Lambert, 2021b). Consequently, Leicester has a much lower historic industrial density (average = 0.59 industries per km$^2$) compared with Glasgow or Belfast (Table 9.5D), and these mainly included light industries. This may be why average soil Pb is higher around light historic industry (LHI), compared with heavy historic industry (HHI) in Leicester (4.41 versus 4.38 log Pb; Figure 9.10C). The opposite pattern was observed in Glasgow/Belfast, likely due to more HHI in these cities (Figure 9.10C).

Despite differences in the industrial pasts of Belfast, Glasgow, and Leicester, historic industry density is consistently the 2nd most important covariate and has a similar positive relationship with soil Pb in all the RF models (Figures 9.9 and 9.10B). Similar results were found in Dublin (Glennon et al., 2014). This may be because it is very difficult to model soil Pb from a single point or historic industrial source. For example, a factory in one city may generate observable soil Pb pollution, but due to local differences or overprinting from other sources, a similar type of factory in another city may not generate observable Pb pollution. The complex nature of urban soils makes it challenging to model soil Pb levels based on historic industry distance, type, and age alone. Alternatively, historic industry density simplifies the relationship between historic industry and soil Pb. It is easier to predict that high soil Pb will occur in areas with lots of industry, as there will have been more diffuse Pb pollution from all the industries, and it is more likely that one of the industries will have been a point source. Furthermore, certain areas of the city such as the city centre and dockyards, tends to contain a high density of historic industry, and this may coincide with the location of other Pb sources e.g. older houses and/or roads.

Although historic industry density correlates with soil Pb in this study and Glennon et al. (2014) there are some issues associated with this method. Firstly, the centroid of each historic industry polygon was input to generate the KDE (in this study) and the SKDE (Glennon et al., 2014). However, this will underrepresent aerially or linearly expansive industries. For example, in the kernel density estimations a small
smithy is equally important as a mile long stretch of railway, or a large chemical factory. Additionally, a KDE requires user input in deciding the search bandwidth and other parameters, which can greatly impact the resulting density estimations.

Land use is a relatively important covariate in the Leicester (3rd most influential) and the Belfast (4th most influential) models (Figure 9.9). Although land use is ranked 6th in the Glasgow model (Figure 9.9A), it shows a significant correlation with soil Pb based on geostatistical analysis (Table 8.5 and Figure 8.12). The PDP plots indicate that there is a greater difference in average soil Pb levels between different land uses in Belfast and Leicester, compared with Glasgow (Figure 9.10D) and is why land use is a more important covariate in these cities. This may be due to less urban regeneration, and/or lower atmospheric Pb pollution in Belfast/Leicester, meaning other Pb sources are clearer. There are also a similar ranking of land uses with regards to soil Pb for all the cities. Specifically, gardens and open space tend to have the highest average soil Pb levels and natural land and road verges the lowest (with some exceptions; Figure 9.10D). However, statistical analysis is required to test if these differences are significant for the Belfast and Leicester datasets. This is similar to other studies e.g. Gibson and Farmer (1983) and Lark and Scheib (2013) (Table 6.2), and results from Paisley and Bishopbriggs (Figure 6.23). Furthermore, garden soil Pb concentrations are practically high in Leicester (Figure 9.10D), perhaps due to sewage sludge amendments, as noted by Bridges (1991).

It is important to note that these covariates may not be applicable to all cities, as these conclusions are only based on three cities. For example, new towns where development was predominantly post-war e.g. Stevenage or Livingston, may not have such high levels of soil Pb due to the lack of pre-1920 buildings and industries. Alternatively, in heavily industrial towns such as Sheffield, the historic industry covariates will be more dominant. Therefore, more investigation is required. Overall, the same drivers of elevated soil Pb concentrations apply across different cities, and this might indicate that it is possible to predict where the high Pb values for each city will occur. However, because of the complex interactions between these drivers and the local context in which they occur it is not possible to predict the magnitude of these highs.
9.6. Conclusions

Lead concentrations in urban soils are often difficult to predict due to lots of over-printing from multiple Pb sources. Predicting areas at high risk from high soil Pb in urban areas is crucial for future land use planning and management. However, this is extremely tricky. This is because hotspots of soil Pb concentrations are often dependent on localised factors that do not consistently generate high Pb concentrations, especially in other cities. Overall, it is best to err on the side of caution and over-predict soil Pb concentrations, rather than underpredict them.

The neighbourhood scale RF models are highly accurate when locating soils with high Pb concentrations, using the 95th prediction percentile or QRF to find the probability of exceeding a threshold (Table 8.7). However, they have inconsistent results when applied to other neighbourhoods (Table 9.6), as they are often over-fit to localised conditions.

The city scale QRF model tends to be more accurate in different cities (Table 9.6). This is likely because the relationship between soil Pb and the covariates is more generalised, due to more samples over a wider area. Furthermore, when applying a model to other cities the probability of high-risk locations is a better method than actual soil Pb predictions, as the model is better at predicting high locations rather than the magnitude of the high values.

When the influence of covariates on soil Pb were investigated at two additional UK cities, a similar pattern in covariate importance was observed. Specifically, OM, building age, historic industry density, and land use, are important covariates when predicting soil Pb in all cities (examined in this study). This suggests that city location only has a minor influence on covariate importance. However, these findings are based on three cities, which have relatively similar developmental histories. Therefore, further investigation is required.
Table 9.6: Summary of external validation statistics for Glasgow, Paisley, and Bishopbriggs RF models. Including statistics of QRF predictions of soil Pb concentrations (all samples show median predictions, high/low show results from the ‘best’ quantile), and high-risk probability prediction accuracy from a confusion matrix (* for high Pb groups only).

<table>
<thead>
<tr>
<th>RF model</th>
<th>Validation dataset</th>
<th>Soil Pb group</th>
<th>Prediction quantile</th>
<th>Mean ME</th>
<th>Mean RMSE</th>
<th>CCC</th>
<th>Pearson’s R²</th>
<th>Overall acc.</th>
<th>Producer’s acc.*</th>
<th>User’s acc.*</th>
<th>Cut off</th>
</tr>
</thead>
<tbody>
<tr>
<td>Greater Glasgow G-BASE</td>
<td>Belfast</td>
<td>All</td>
<td>50&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.5</td>
<td>0.74</td>
<td>0.13</td>
<td>0.27</td>
<td>71%</td>
<td>53%</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>95&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.33</td>
<td>0.59</td>
<td>0.08</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.33</td>
<td>0.52</td>
<td>0.14</td>
<td>0.24</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Leicester</td>
<td>All</td>
<td>50&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.49</td>
<td>0.72</td>
<td>0.18</td>
<td>0.46</td>
<td>80%</td>
<td>47%</td>
<td>29%</td>
<td>&gt;30%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>75&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.35</td>
<td>0.54</td>
<td>0.08</td>
<td>0.11</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.33</td>
<td>0.5</td>
<td>0.17</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Tamar</td>
<td>All</td>
<td>50&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1.17</td>
<td>1.2</td>
<td>0.03</td>
<td>0.29</td>
<td>69%</td>
<td>67%</td>
<td>4%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>High</td>
<td>75&lt;sup&gt;th&lt;/sup&gt;</td>
<td>-0.35</td>
<td>0.39</td>
<td>0.09</td>
<td>0.21</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.87</td>
<td>0.9</td>
<td>0.04</td>
<td>0.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Paisley Bishopbriggs</td>
<td>Bishopbriggs</td>
<td>All</td>
<td>50&lt;sup&gt;th&lt;/sup&gt;</td>
<td>-0.87</td>
<td>0.91</td>
<td>0.04</td>
<td>0.27</td>
<td>53%</td>
<td>12%</td>
<td>33%</td>
<td>&gt;30%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Low</td>
<td>25&lt;sup&gt;th&lt;/sup&gt;</td>
<td>0.95</td>
<td>1.11</td>
<td>0.21</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bishopbriggs</td>
<td>Paisley</td>
<td>All</td>
<td>50&lt;sup&gt;th&lt;/sup&gt;</td>
<td>1.44</td>
<td>1.48</td>
<td>0.12</td>
<td>0.39</td>
<td>44%</td>
<td>100%</td>
<td>42%</td>
<td>&gt;40%</td>
</tr>
</tbody>
</table>
**References - Chapter 9**


Appendix 9.1 - Example validation Matlab code

%1. Load dataset with trained RF model
load G_BASE5

%2. Load validation dataset
load Leister2; % (656 points)
len = length(Pb);
% Rename each covariate to match covariates in Glasgow model
log_OM = log(OM);
buildingage = BA;
HI_density_log = log(hist_ind_desnity);
IND_date2 = IND_date2;
IND_type = IND_type;
% Put variables in table
T2p = table (log_OM, buildingage, HI_density_log, IND_date2, LandUse, IND_type);

%3. Risk probability calculation and validation
P_vals = 0.01:0.01:0.99; % define percentiles
Lp = length(P_vals);
pred2 = quantilePredict(rf_L, T2p, 'Quantile', P_vals); % run QRF
threshold = 5.2983; % define high soil Pb threshold (200 mg kg)
prob = 100*sum(pred2>threshold,2)/(Lp+1); % calculate probability
hist(prob) % plot histogram of the probabilities
title('Probability soil Pb exceeds threshold - Leicester')
xlabel('Probability soil Pb >200mg/kg (%)')
root = sum(sqrt(prob));
txt = {'Sum Sq Root Prob: ' root};
text(50,100,txt);

% Check if samples predicted to have high soil Pb match actual Pb
Pb_cat = ones(len,1); % classify observed Pb values
Pb_cat = categorical(Pb_cat);
Pb_cat(Pb>=200) = 'High';
Pb_cat(Pb<200) = 'Low';
valueset = {'High' 'Low'}; % remove empty 1 class
Pb_cat = categorical(Pb_cat,valueset);
summary(Pb_cat)

prob2 = categorical(prob); % classify predicted Pb values
prob2(prob>=30) = 'High';
prob2(prob<30) = 'Low';
valueset = {'High' 'Low'}; % remove empty 1 class
prob2=categorical(prob2,valueset);
summary(prob2)
plotconfusion(Pb_cat,prob2) % compare using a confusion matrix

% Output as table with coordinates so can map in GIS
output = table(EASTING,NORTHING,prob,prob2);
writetable(output, 'Predicted_prob_Leicester_30per.csv');

%4. Soil Pb concentration prediction and validation
pred1 = quantilePredict(rf_L, T2p, 'Quantile', [0.25 0.5 0.95]); % run QRF
% Compare predicted and actual values
actual=(log(Pb));
% All values validation
Valresults = ones(3,6); % save validation results in table
for i = 1 : 3
    me_rf = pred1(:,i)-actual; %mean error
    valresults(i,1) = mean(me_rf); %save in column 1
    rmse_rf = mean(sqrt((pred1(:,i)-actual).^2)); %RMSE
    valresults(i,2) = mean(rmse_rf); %save in column 2
    % Calculate the concordance correlation coefficient
    v1 = nanvar(pred1(:,i));
    v2 = nanvar(actual);
    m1 = nanmean(pred1(:,i));
    m2 = nanmean(actual);
    p = corr(pred1(:,i),actual,'rows','pairwise');
    r2_c = 2*sqrt(v1)*sqrt(v2)*p/(v1+v2+(m1-m2)^2); %save in column 3
    % Calculate Pearson’s correlation coefficient
    r2 = corrcoef(pred1(:,i),actual);
    valresults(i,4) = r2(2,1); %save in column 4
    valresults(i,5) = min(pred1(:,i)); %save min vale in column 5
    valresults(i,6) = max(pred1(:,i)); %save max vale in column 6
end
round(valresults,2)
% Assess accuracy of each soil Pb group i.e. high, and low
T1 = table(actual, pred1); %combine actual and predicted values
high = T1((T1.actual>=5.3),:); %select high samples
low = T1((T1.actual<5.3),:); %select low samples
group=high; %change group to get different stats
group_ar = table2array(group);
group_actual = group_ar(:,1); % select 1st column (actual Pb values)
group_pred = group_ar(:,[2:4]); %select columns 2-7 (predicted Pb)
% Run validation stats
valresults_g = ones(3,6);
for i = 1 : 6
    round(valresults_g,2)
end
% Plot results coloured by soil Pb group
low_ar =table2array(low);
low_actual = low_ar(:,1);
low_pred = low_ar(:,[2:4]);
high_ar = table2array(high);
high_actual = high_ar(:,1);
high_pred = high_ar(:,[2:4]);
t = tiledlayout(1,3,'TileSpacing','compact');
titles = {'25th', '50th', '95th'};
x = [2.5 9.5]; % fix x and y axis to match
y = [2.5 9.5];
hold 'off';
for i = 1 : 3
    nexttile
    line(x,y,'Color','k','LineStyle','--');
    hold 'on';
s1=scatter(low_actual,low_pred(:,i),9,'c','filled','MarkerEdgeColor','k');
s3=scatter(high_actual,high_pred(:,i),9,'m','filled','MarkerEdgeColor','k');
title(titles(i));
xlabel('Actual soil Pb (log)');
ylabel('Predicted soil Pb (log)')
end
legend([s1, s3], 'Low', 'High');
set(legend, 'Location', 'southeast');

%5. Make own RF to assess predictors
z = log(Pb);
mns = 6;
rf_L = TreeBagger(1001, T2p, z, 'Method', 'regression',
'OOBPermutedPredictorDeltaError', 'on', 'PredictorSelection', 'curvature',
'InBagFraction', 0.5,'NumPredictorsToSample', 4, 'MaxNumSplits',
mns);

% Extract and plot the importance of each covariate
imp = rf_L.OOBPermutedPredictorDeltaError;
figure;
bplot = bar(imp);
bplot.FaceColor = '#808080';
importance = round(imp,2);
ylabel('Predictor importance estimates');
xlabel('Predictors');
title('Leicester');
text(1:length(importance),importance,num2str(importance'),
'vert', 'bottom', 'horiz', 'center');
xticklabels({'OM (log)', 'Build age', 'Ind density (log)', 'Ind age',
'Land use', 'Ind type'})
xtickangle(45);

% Create PDP - relationship between soil Pb and a covariate
figure
t = tiledlayout(3,2,'TileSpacing','compact');
title(t, 'PDP')
for i = 1 : 6
nexttile
plotPartialDependence(rf_L,i)
title('')
ylabel 'Soil Pb (log)'
end

%6. Risk predictions at prediction grid
load grid_L %import grid with covariates, regroup to match Glasgow
T2p = table(log_OM, buildingage, HI_density_log, IND_date2, LandUse,
IND_type);

% Predict at grid
P_vals = 0.01:0.01:0.99;
Lp = length(P_vals);
pred2 = quantilePredict(rf2,T2p,'Quantile',P_vals);
threshold = 5.2983;
prob = 100*sum(pred2>threshold,2)/(Lp+1);

% Plot location of probabilities
X=[x y];
figure(1)
scatter(X(:,1),X(:,2),100,prob,'filled')
cb = colorbar; % creates the colorbar on side

% Output so can map in GIS
output = table(x, y, prob,pred2);
writetable(output, 'Grid_prob_L.csv');
Chapter 10. Discussion and conclusions.
This final chapter aims to consider how new techniques were used to achieve the original objective of the project. Additionally, the key findings from each individual chapter will be discussed. The over-arching themes from the entire thesis will then be explored, such as the scalability and transferability of the models, comparison of techniques, and communication of methods. These relate to the research questions posed in Section 1.2. Lastly, further avenues of research relating to this thesis will be identified.

10.1. Initial aim and novel methods

The original intention of this thesis was to design a model which could accurately predict soil Pb concentrations in a British urban environment. This would build on the existing international literature (e.g. Schwarz et al. (2013) and references in Table 8.1) but focuses on UK cities and uses a geostatistical approach to select covariates. Subsequent to this aim being established, similar work has been conducted by Wragg and Cave (2021), which uses a unique methodology to demonstrate how machine learning modelling of soil Pb can be used as part of a land contamination assessment. This underscores the value of more localised mapping of soil Pb concentrations, to help decision makers pinpoint locations most at-risk from soil Pb in a city. These locations will require further investigation but will ultimately save time and money in any future remediation project. Therefore, the model generated by the research in this thesis could be used as a starting point, from which the potential health impacts from soil Pb can be assessed and mitigated against.

Overall, this thesis has proven that a machine learning model, i.e. QRF, can estimate soil Pb concentrations in a British urban environment. Furthermore, this model explained a substantial proportion of the variability of Pb (e.g. in the Glasgow training datasets Pearson's R² > 0.5 and overall accuracy > 74%; Table 8.7). In line with previous studies, the model was used to predict the expected (median or 50th percentile) value of soil Pb at each location. Generally, land managers and decision makers are more concerned about the risk of high Pb concentrations than the variation of the median value of the predictions. Therefore, this project focusses on the risk of these high values by also considering the 95th percentiles of the QRF predictions.
One potential weakness of machine learning methods, including QRFs, is that they can identify patterns in the calibration data that have occurred at random, rather than reflecting the key processes driving variation in soil Pb concentrations. In these circumstances the patterns are unlikely to appear in other datasets and the model is said to be overfitted. To overcome this issue, this project used LMM and Wald tests to test hypotheses regarding the causes of soil-Pb variation in a manner that accounts for spatial dependence of the Pb measurements. The hypotheses were proposed based upon existing knowledge of the drivers of Pb variation and were posed in terms of potential linear relationships between Pb concentrations and relevant covariates. This helped ensure the QRF model was not overfit to the conditions in Glasgow, since covariate terms were only included where there was evidence that the relationships observed in the data agreed with expert understanding of Pb variation.

This project also examines the importance of scale when modelling soil Pb. Most studies which model soil Pb focus on a single scale (see references in Table 8.1), although there are examples of multi-scale modelling. For example, Hooker and Nathanial’s (2006) study of two BGS datasets in Wolverhampton, and Bugdalski et al.’s (2014) analysis of high-scale soil Pb variability in Detroit’s urban gardens. The scale a model is trained on will control the accuracy and useability of any output maps and is largely controlled by the training dataset’s sampling strategy. This research has indicated that soil Pb concentrations have a different set of controlling variables at the city scale, compared with the neighbourhood scale (Tables 6.10 and 6.12). Consequently, different covariates should be input into models at varying scales. The influence of scale is explored in more detail in Section 10.3.1.

Studies often cite that a strength of machine learning models is their transferability i.e. the models can use covariate datasets to estimate soil Pb concentrations in regions or cities that there are no current soil Pb measurements for. However, applying a QRF model to estimate soil Pb in a new city has not been attempted in any other studies, to the authors knowledge. This is likely because it is extremely challenging to do so, as soil Pb concentrations are often controlled by individual, city specific factors. Therefore, another unique part of this study is that the machine learning models were tested in cities they were not trained in (e.g. Belfast and Leicester, see Chapter 9). This ensures that the models have not been over-fit to conditions in one city, e.g. Glasgow, which is vital if the model is to be applied
widely. The Glasgow trained QRF model performed better than expected, when it was used to predict soil Pb concentrations and risk in Belfast, Leicester, and the rural Tamar Catchment (i.e. Pearson’s $R^2 > 0.27$, overall accuracy $> 69%$; Table 9.6). The transferability of QRF models is examined further in Section 10.3.2.

Lastly, the thesis explores effective presentation methods for the results from predictive mapping, in particular, how to convey uncertainty. This is crucial if the results from machine learning models are to have practical use. However, none of the previous studies modelling soil Pb considered here (Table 8.1) accounted for uncertainty in their predictions, apart from Kirkwood et al. (2016) who used QRF to quantify the 95% confidence interval. This study uses QRF to predict soil Pb concentrations and their confidence intervals (e.g. Figures 8.5, 8.14 and 8.15), but also predicts locations of elevated soil Pb risk i.e. the probability that a location will have high soil Pb concentrations ($\geq 200$ mg kg$^{-1}$; e.g. Figures 8.8 and 8.18). This is discussed in more depth in Section 10.3.4.

10.2. Summary of main findings

10.2.1. Exploratory analysis of soil Pb concentrations

Soil Pb concentrations are found to be enhanced above their natural background levels in all three Glasgow datasets – Greater Glasgow, Paisley, and Bishopbriggs. Furthermore, all datasets show positive skew due to a few high outliers. A Box-Cox transform generates a more Normal distribution of values, which is required for many types of further statistical analysis. Lastly, the soil Pb concentrations are generally isotropic, so no strong directional trends need to be corrected for in future modelling.

10.2.2. The spatial dependence and variability of soil Pb concentrations

Spatial correlation or dependence refers to proximal measurements being more likely to be similar than those made a large distance apart. Understanding spatial dependence and variability is vital when answering the first research question (Section 1.2): how important is scale when modelling and mapping soil Pb concentrations?

This study found that soil Pb concentrations are spatially dependent, as proven by the positive and significant Moran’s I values in all study areas (Global Moran’s I = $+0.15$ to $+0.33$, $p<<0.01$; Table 5.3). This indicated that future statistical analyses
should account for spatial dependence. Although the total amount of spatial dependence did not vary with scale (i.e. Bishopbriggs and Greater Glasgow had similar Moran’s I values), LISA measurements demonstrated that spatial dependence was heterogenous across the study areas, and tended to cluster in hotspots and coldspots (Figure 5.4).

The distance over which spatial dependence persisted (i.e. the range of spatial dependence) was explored in each study area using REML-fit semivariograms and ordinary kriging (OK). The range was controlled by the sampling scale, as the Greater Glasgow semivariogram had a 100-fold higher effective range, compared with the range of the neighbourhood semivariograms (10,506 m vs 72 or 116 m; Table 5.5 and Figure 5.5). This result is expected, as variation of soil Pb concentrations due to large scale processes such as diffuse pollution or geology are likely to be apparent in the Greater Glasgow data, whereas these same factors are likely to be relatively constant within the neighbourhood surveys. In the neighbourhood datasets other factors, such as land use, are likely to be more influential. The OK maps revealed spatial patterns in soil Pb concentrations at all scales (Figures 5.6 to 5.8). These hotspots and coldspots of soil Pb concentrations were likely caused by a combination of anthropogenic and environmental factors.

10.2.3. Sources of soil Pb

To determine what could be causing the spatial heterogeneity in soil Pb concentrations, the literature was examined. Other studies (Tables 6.1 to 6.5) have found that soil Pb concentrations can be controlled by 12 key covariates, namely: land use, road distance, traffic density, historic industry (type, distance, age, and density), building age and distance, and soil factors (OM content, pH, and soil texture). This knowledge assists in answering the second research question (Section 1.2):

“Can the sources of Pb in an urban environment be ascertained by exploring the relationships between sampled soil Pb chemistry and landscape features?”

To address this question the relationship between soil Pb concentrations and each covariate were explored using model-based geostatistics, as spatial dependence is present in all soil Pb datasets. Specifically, the REML-fit semivariograms were extended to include covariates (Equation 5.8). Furthermore, the Wald statistic (Equation 6.4) determined if the covariate had a significant impact on soil Pb
concentrations in each study area, and also quantified the strength of the relationship. Each covariate was included when mapping soil Pb concentrations using the E-BLUP.

At the city scale, six covariates had a significant relationship with soil Pb concentrations (Table 6.10). Organic matter had the strongest relationship with soil Pb (Wald = 507), then building age (Wald = 42), historic industry density (Wald = 21), soil texture (Wald = 13), history industry age (Wald = 12), and road type (Wald = 4). However, soil Pb's relationship with soil texture and road type did not accord with previous studies, perhaps due to outliers in some small groups, the unique conditions in Glasgow, or there being cofounding relationship between these factors and other drivers of Pb variation. Therefore, soil texture and road type should not be included as covariates in future predictive models, as this would risk over-fitting the model to the conditions in Glasgow. Furthermore, land use and historic industry type did not have a significant relationship with soil Pb (p = 0.055 and 0.057). However, the Wald values indicated a relatively strong relationship between these covariates and soil Pb (Wald = 11 and 4 respectively; Table 6.10), and the relationships agreed with the literature. Consequently, land use and historic industry type could be recommended for inclusion in future models for Glasgow city.

At the neighbourhood scale, OM was also found to be the most influential covariate, followed by land use, road and building distance, and building age (Table 6.12). Despite some covariates showing a significant relationship with soil Pb concentrations in both study areas e.g. OM and land use, the influence of other covariates varied due to local factors. For example, possible reworking of contaminated soil during recent house construction (e.g. in Paisley, Figure 6.30), and a historic refuse heap contaminating soil far from roads and houses (e.g. in Bishopbriggs, Figure 6.29). This suggests that neighbourhood scale machine learning models may have limited transferability.

10.2.4. The origins of soil Pb in Paisley and Bishopbriggs

The previous section described how model-based geostatistics can be used to determine the sources of soil Pb in an urban environment. However, some potential Pb sources, such as road distance and historic industry distance, did not always show a significant relationship with soil Pb (Figures 6.13 and 6.15), despite the general negative relationship proven in other studies (see Tables 6.3 and 6.5). This
may be due to geographical and historical “over-printing” from multiple soil Pb sources in these old urban areas, making the relationship between soil Pb concentrations and a single source difficult to untangle. Lead isotopes offer a potential solution to this issue, as the \(^{206}\text{Pb}/^{207}\text{Pb}\) and \(^{208}\text{Pb}/^{207}\text{Pb}\) ratios in the soil reflect the original isotopic ratios in the parent ore. This can be used to apportion Pb originating from petrol and industrial/coal sources.

Consequently, Pb isotopes in half the soil samples collected from Paisley and Bishopbriggs were determined using ICP-MS. The Pb isotopic ratios were then used to determine the percentage and total concentration of Pb originating from petrol and coal sources (Equation 7.1, Table 7.2). These were modelled and mapped in Paisley and Bishopbriggs using REML-fit semivariograms and OK. Furthermore, the influence of covariates, i.e. the distance to the nearest road and historic industry location, were modelled using the same methodology as in Chapter 6.

The analysis showed that most soil Pb originated from coal sources in both Paisley and Bishopbriggs (mean = 66% (316 mg kg\(^{-1}\)) and 78% (340 mg kg\(^{-1}\)) respectively, Table 7.3). This Pb may have entered the soil via coal mining emissions and waste, but the majority of coal Pb will likely have been deposited to the soil from background emissions e.g. coal burning as a power or heat source in industrial and domestic settings. Dumping of ash waste to fertilise garden soils may also account for some Pb present in domestic garden soils.

Soil in Paisley has a higher percentage and concentration of Pb from petrol, compared with Bishopbriggs (34% vs 22% and 199 mg kg\(^{-1}\) vs 78 mg kg\(^{-1}\), Table 7.3). This is possibly due to differences in traffic density, and roadside soil conditions influencing soil Pb mobility, but requires further investigation.

When mapped, petrol soil Pb hotspots broadly correlate with the locations of roads, and high values of coal Pb correspond with historic industry locations (Figures 7.5 and 7.6). However, this pattern is not consistent, indicating other factors are influencing the distribution of soil Pb from petrol or coal and/or their isotopic signature e.g. general background industrial emissions, older houses burning coal fires and disposing of ash on garden soil, and the use of leaded paint. Furthermore, the isotopes only relate to the topsoil which may have been disturbed by more recent activity e.g. house building. Soil properties, and subsequently Pb mobility, are also important but have not been accounted for.
There is a significant relationship between total Pb, petrol Pb, and coal Pb concentrations, with distance to a road. These relationships are much stronger in Bishopbriggs compared with Paisley. However, road distance and petrol Pb have an unexpectedly positive correlation in Bishopbriggs, which seems to be largely driven by a refuse heap far from roads generating soil Pb pollution. The road distance and petrol soil Pb trend is negative in Paisley, as historic leaded-petrol emissions are expected to have elevated Pb levels in roadside soil (Table 6.3).

There is no relationship between historic industry distance and coal, petrol, and total Pb in either location, which agrees with earlier research (Chapter 6). This may be because coal Pb pollution was so widespread as background emissions, that it was not limited to specific locations or historic industries. Furthermore, not all types of historic industry will have generated Pb pollution, so the examples of historic industry in Paisley and Bishopbriggs may be too specific.

### 10.2.5. Machine learning prediction of soil Pb

Identification of the sources of soil Pb in Chapters 6 and 7 was used to decide which covariates should be to input into machine learning models. This knowledge is needed to address part of the third, and arguably most important, research question of the thesis:

“Are models (both geostatistical and machine learning) a viable method for accurately predicting soil Pb concentrations?”

Random forest models were built for each Glasgow dataset i.e. the Greater Glasgow city dataset, and the two neighbourhood datasets - Paisley and Bishopbriggs. These machine learning models were used to predict and map soil Pb concentrations at the city and neighbourhood scales, and their accuracy was assessed using 10-fold cross-validation.

The map estimating soil Pb concentrations at the city scale (e.g. Figure 8.4) showed several hotspots, most notably over the East End and in Paisley. This somewhat agreed with the kriged soil Pb map (e.g. Figure 5.6A). The predictions were reasonably accurate as the Pearson’s correlation coefficient between predicted and sampled values at the 50th percentile was 0.51, and the CCC was 0.33 (Table 8.5). These validation statistics show an improvement, compared with results from other studies using RF models to predict soil Pb concentrations with similar Pb maximum levels (Table 8.1). However, the predicted span of soil Pb concentrations in this
study is much narrower than the actual span of soil Pb values (Table 8.5, Figure 8.7), due to the central tendency issue with RF (i.e. tendency for predictions to be smoother than observed variation; Zhang and Lu, 2012).

Quantile regression forests allowed the examination of soil Pb predictions at different percentiles. Furthermore, the predictions were split by their soil Pb concentrations, with high samples containing ≥ 200 mg kg\(^{-1}\) of soil Pb, and low samples containing less than this threshold. Unsurprisingly, higher percentiles (i.e. the 75\(^{th}\) or 95\(^{th}\) percentile) more accurately predicted high soil Pb concentrations, and vice versa for lower (i.e. 5\(^{th}\) or 25\(^{th}\)) percentiles (Table 8.5 and Figure 8.7). This is useful knowledge for future discovery of soil Pb hotspots, which pose the greatest potential health risk.

Predicted hotspots and coldspots of soil Pb concentrations were also observed at the neighbourhood scales (e.g. Figure 8.13). These generally corresponded to kriged soil Pb concentrations (e.g. Figures 5.6B and C) but tended to be less smooth due to the influence of geographic covariates, such as roads and land use boundaries. These predictions were more accurate than the city-wide estimations in Paisley (Pearson’s R\(^2\) = 0.62, CCC = 0.56; Table 8.6A) but not always in Bishopbriggs (R\(^2\) = 0.59, CCC = 0.35; Table 8.6B), perhaps due to the range of soil Pb values and outliers. However, the 75\(^{th}\) percentile predictions of only samples with high soil Pb concentrations were very accurate in Bishopbriggs (R\(^2\) = 0.75, CCC = 0.77; Table 8.6B, Figure 8.17). Overall, the neighbourhood scale models were more accurate when predicting high soil Pb locations, compared with the city scale model.

Chapter 8 also answers research question five, i.e. can the machine learning models improve our understanding of areas that present a greater risk from soil Pb? This question was initially investigated in Section 8.4.1.2. using high percentile predictions of soil Pb concentrations. However, it was explored further using QRF to find the quantile where a high soil Pb threshold (200 mg kg\(^{-1}\)) was exceeded. This was the probability that a location will have high soil Pb concentrations, and these probabilities were mapped for each study area.

The predicted risk maps for Glasgow (Figure 8.8) and Paisley and Bishopbriggs (Figure 8.18) show similar patterns to the predicted soil Pb concentration maps. Locations with high soil Pb risk are emphasised by adding contour lines around hotspots with a > 95 or 99% significance, based on Getis-Ord Gi* cluster analysis.
The accuracy of the probability maps was assessed by comparing the predicted group (high = ≥ 30% probability of containing ≥ 200 mg kg\(^{-1}\) of soil Pb; low = < 30% probability of high soil Pb, based on ROC analysis) with observed groups (high = ≥ 200 mg kg\(^{-1}\) of soil Pb, low < 200 mg kg\(^{-1}\)) in a confusion matrix. The Glasgow risk model was highly accurate (74% overall accuracy) and had good producer’s and user’s accuracies of high samples (68% and 53% respectively; Figure 8.10). The neighbourhood scale models had even higher overall accuracies (84% and 83% for Paisley and Bishopbriggs), and producer’s and user’s accuracies > 70% (Figure 8.20). This indicated that the majority of locations with high soil Pb concentrations were classified as high risk.

### 10.2.6. Modelling in other cities

Cross-validation statistics are a valid method to assess model accuracies; however, this type of analysis does not test how well the machine learning models preform in other cities or even in rural regions. Subsequently, research question four asks if machine learning models trained in one city are applicable to other cities or geographical locations? (Section 1.2). This is an important, and often over-looked, issue because the main strength of machine learning models is that they can be used to help screen soil Pb concentrations or risk, in regions where a soil Pb dataset does not exist or is too sparsely sampled to generate an accurate map (Section 1.1). This question was explored in Chapter 9, where three additional G-BASE (or Tellus) datasets (i.e. for Belfast, Leicester, and the Tamar catchment) were used to validate the Glasgow QRF model.

Overall, the Glasgow city-scale machine learning model explained some variability when used to predict soil Pb concentrations in Belfast, Leicester, and even the more rural Tamar catchment (all samples median prediction Pearson’s \(R^2 = 0.27 \sim 0.47\); CCC = 0.03 \sim 0.18; Table 9.2). The relatively small concordance correlations are likely to be an indication that the test cities have different mean soil Pb concentrations to Glasgow, but the higher Pearson correlations indicate that similar factors in the different cities are leading to Pb hotspots and coldspots. This is further demonstrated by the much narrower range in the predicted soil Pb concentrations, compared with the actual soil Pb concentrations (Figure 9.3). This was likely due to the central tendency issue with RF, which was also observed when cross-validating the Glasgow model.
The Glasgow model accuracy was slightly lower when only samples with high soil Pb concentrations (≥ 200 mg kg⁻¹) were validated (Pearson’s $R^2 = 0.1 - 0.18$; CCC = 0.06 - 0.08; Table 9.2). This may be because Glasgow has higher average soil Pb concentrations, likely from more intense industrial background emissions, compared with the other cities. Subsequently, the Glasgow model may over-predict background soil Pb concentrations in other cities but miss some soil Pb hotspots. This could be accounted for by using a higher percentile e.g. the 99th percentile, when predicting high locations. However, the fact that the CCC is positive in all other geographical locations indicates that the Glasgow model explains some variation in other cities and regions, which is difficult to achieve.

An interesting finding was that the Glasgow city model performed much better when it was used to specifically predict locations at-risk (≥ 30% probability) of having high soil Pb concentrations (≥ 200 mg kg⁻¹) in other cities. The overall accuracy of the confusion matrices were 69% to 80%, with the producer’s accuracy of high samples > 47% (Table 9.3). This indicates that almost half the samples with high soil Pb were predicted to be high-risk in other cities. Overall, the Glasgow RF model can predict locations likely of high soil Pb in other cities; however, it often under-estimates the magnitude of these values but it is still useful as an initial screening method.

The accuracy of the neighbourhood scale models was also examined by using the other neighbourhood dataset to validate predictions. Bishopbriggs predictions using the Paisley RF model were extremely poor (CCC = 0.04, Table 9.4A; high samples producer’s accuracy = 12%, Figure 9.8A). This may occur because soil Pb concentrations tend to be slightly higher in Bishopbriggs, compared with Paisley; therefore, the Paisley model will under-predict soil Pb concentrations and risk. Furthermore, Chapter 6 indicated that covariates have a varying relationship with soil Pb, depending on the neighbourhood examined. Consequently, the model may be over-fit to local conditions.

On the other hand, when the Bishopbriggs RF model was used to predict soil Pb concentrations and risk in Paisley the results were reasonable (CCC = 0.12, Table 9.4B; high samples producer’s accuracy = 100%, Figure 9.8B). This is likely because the higher soil Pb values in Bishopbriggs means the RF model trained in that neighbourhood over-predicts soil Pb in Paisley. This is also supported as the most accurate prediction was for the lowest percentile (25th, CCC = 0.21; Table 9.4B).
10.3. **General themes**

Although each chapter provides specific answers to some of the research questions outlined in Section 1.2, there are also overarching themes to the thesis that require further discussion.

10.3.1. **Scalability: the importance of geographical scale**

Atkinson and Tate (2000) refer to geographical scale as the size or extent a process occurs over, which is composed of scales of spatial measurement and scales of spatial variation. Furthermore, Openshaw and Taylor’s modifiable aerial unit problem (MAUP) discusses how the relationship between a covariate and a fixed variable can change, from a positive to negative correlation, depending on the scale, and level of aggregation, it is examined at (Openshaw and Taylor, 1979).

These issues of scale are explored by addressing research question one:

“How important is scale when mapping and modelling soil Pb concentrations?”

As discussed previously, most studies mapping and modelling soil Pb concentrations focus on a single sampling and mapping scale (e.g. references in Tables 5.1 and 8.1), although there are exceptions. This research benefits from having access to three soil Pb datasets, collected at different sampling scales within the same city: one at the city scale and two at the neighbourhood scale.

When examining the scale of spatial variation, soil Pb concentrations are spatially dependent from 50 m to 10.5 km (based on semivariogram ranges, Table 5.5). This is a large range but agrees with other studies which found the semivariogram range varied between 20 m and 85 km (Table 5.1). Overall, results from this study, and others, indicates that the spatial range of soil Pb concentrations in a semivariogram model is predominantly controlled by the sampling scale and the large-scale processes that can influence the data from a larger study area e.g. background emissions. Therefore, this should be carefully considered, as it can impact future mapping and modelling resolution.

Furthermore, soil Pb variation, witnessed as hotspots and coldspots, can be found at all scales in Glasgow (Figures 5.4 and 5.6). This corresponds with other research, where soil Pb hotspots have been observed at the microscale (e.g. < 10 m; Bugdalski et al., 2014), local scale (e.g. < 1 km; Schwarz et al., 2012), city scale (e.g. < 10 km; Lark and Scheib, 2013), national scale (e.g. < 100 km; Savignan et al.,
2021), and even the continental scale (e.g. < 2,500 km; Reimann et al., 2012). The lack of homogeneity in soil Pb at even the ultra-fine scale proves that size does matter when mapping and/or modelling soil Pb, especially in urban environments. Something that has also been noted by other researchers (e.g. Bugdalski et al., 2014). The scale studied in any piece of work will impact the sampling, modelling, and mapping strategies and resolution, with regards to soil Pb concentrations.

The sampling strategy or measurement scale of any study will depend on the size of the study area, the available budget, and most importantly the project goals. However, Oliver and Webster (2015) suggest that the sampling interval should be approximately half the effective range in order to capture the correlated structure of a variable. Therefore, this study has indicated that soil Pb should be sampled on a scale ~ 25 m if soil Pb variation is to be sufficiently mapped at the neighbourhood scale.

The sampling scale can also impact the mapping resolution of soil Pb. For example, Figures 5.9 and 5.10 illustrate the differences in ordinary kriged maps for the same area (750 x 300 m) using the city dataset with a 200 m sampling interval, and the local datasets with a 30 m mean sample separation distance. Much more detail can be observed on the maps generated from the neighbourhood datasets, compared with the city dataset. Overall, it is inevitable that some soil Pb hotspots will be overlooked at a higher sampling scale. However, larger scale maps have the advantage of mapping soil Pb over a wider area.

The level of mapping detail ultimately depends on the needs of the end user. For example, a local authority planning future land use needs to understand soil Pb variability at the city scale, whereas an individual homeowner would want to know the dangers posed by soil Pb in their garden. As this study has proven, soil Pb can be heterogeneous over a short distance (i.e. < 50 m); therefore, the sampling scale needs to be carefully considered before mapping and predictive modelling.

Scale can influence soil Pb machine learning models by dictating which covariates are selected, as the relationship between soil Pb concentrations and the covariates varies over the 50 m to 10.5 km scale. Section 6.4.4. discusses the importance of scale with regards to soil Pb’s relationship with each covariate and concludes that the relationship between some covariates and soil Pb concentrations is impacted by scale. For example, road and building distance appear more influential on soil Pb concentrations in the neighbourhood datasets, compared with the city-wide
dataset, due to difference in sampling strategies (Tables 6.10 and 6.12). Conversely, building age and historic industry benefit from the more numerous examples in the city dataset, which down weights the influence of any outliers. Other covariates such as the soil OM content have a similar influence on soil Pb concentrations regardless of scale. Although, Section 6.4.3. proves that the relationship between these covariates and soil Pb are also influenced by location. Ideally, a combination of sampling scales would be required to accurately account for the influence of covariates on Pb in heterogenous urban soils. This fits with Openshaw and Taylor’s MAUP and warns other researchers to be wary of scale when choosing covariates for predictive models, as one scale does not always fit all.

The machine learning model accuracy can also be impacted by scale. For instance, the neighbourhood scale models tend to be slightly more accurate than the city model (Table 8.7). This is evident when looking at the overall accuracy of the confusion matrix, which assessed how well the QRF models predicted high risk areas (neighbourhood models = 84%, city model = 74% overall accuracy; Table 8.7). The neighbourhood models also have higher cross-validation accuracy when high quantiles (e.g. the 95\textsuperscript{th}) from QRF are used to predict high soil Pb concentrations (CCC = 0.49 to 0.75 for neighbourhood models, CCC = 0.41 for the city model; Table 8.7). However, the neighbourhood models are less accurate when applied to other geographical locations (Table 9.4), as discussed in the next section.

10.3.2. Transferability: the importance of geographical location

The most successful machine learning models are transferable i.e. they should be able to identify the most at-risk locations from high soil Pb concentrations in other similar cities or regions. However, as discussed previously, the transferability of soil Pb QRF models have not been tested using new datasets from cities/regions the model was not trained in.

The transferability of RFs has been examined in other research fields, with varying success. For example, Jin et al. (2018) proved than a RF model predicting canopy height from remotely sensed data was only transferable, when it was trained using samples from various regions and vegetation types. Similarly, Orynbaikzy et al. (2022) tested the transferability of a RF model predicting large-scale crop type across Germany and found that covariates had to be carefully selected. Lastly,
Bjerre et al. (2022) conducted a spatial cross-validation study on nine sub-basins to determine the transferability of a RF model predicting groundwater drainage. They observed a decrease in validation statistics when the model was applied to new regions ($R^2 = 0.79$ in training region, $R^2 = 0.13$ to 0.61 in new regions).

There tends to be a trade-off between model accuracy and model transferability, with more covariates improving internal cross-validation statistics but reducing model accuracy when it is transferred to a new geographical location. This can be observed by comparing the results from other research. For example in Table 8.1 some studies use 20 plus covariates to train their RF model and have very good validation statistics, but crucially do not test their model in other geographical locations.

The number of covariates included in a RF model was briefly explored in the Glasgow RF model, where extra covariates were found to improve the cross-validation statistics but risked over-fitting the model to the specific conditions in Glasgow. For instance, geostatistical analysis indicated that soil texture and soil Pb concentrations had a strong relationship in the Greater Glasgow dataset (Wald = 13, $p <<0.01$; Table 6.10). Furthermore, including soil texture in the RF model improved the Pearson’s $R^2$ value by 0.18 (CCC by 0.25). However, sandy and silty soils had the highest soil Pb concentrations (median soil Pb (Box-Cox) = 2.93), with clay rich soils containing less Pb (median soil Pb (Box-Cox) = 2.88; Figure 6.21). This disagrees with the literature, where clayey soils tend to contain the most Pb, due to their higher cation exchange capacity and less porous nature reducing Pb mobility (Walraven et al., 2013). The opposite relationship in Glasgow is likely due to sandy or silty alluvial soils occurring near the River Clyde which runs through the centre of Glasgow. The city centre also has the highest soil Pb concentrations from other sources e.g. petrol, paint, and historic industry. Therefore, including soil texture in the Glasgow RF model slightly reduced accuracy in the Belfast and Leicester predictions, where different environmental factors were found to influence soil Pb concentrations ($R^2$ decrease of 0.1 and CCC decrease of 0.2). If numerous over-fitted covariates had been included in the Glasgow RF model, then the reduction in accuracy may have been more significant. Overall, the value of each covariate should be considered before it is input in a RF model.

The transferability of a RF model is not only dependant on the number of covariates, but also on how consistent their relationship is with soil Pb in other cities. For
instance, the relationship between soil Pb and some covariates can be influenced by geographical location, as discussed in Section 9.5.2. Including covariates which are known to be influenced by local factors will reduce the transferability of the model. For example, OM is considerably less important in Belfast (ranked 3rd) and Leicester (ranked 4th), compared with Glasgow (ranked clearly 1st; Figure 9.9). This is likely caused by higher historic Pb atmospheric deposition rates in Glasgow, compared with Leicester and Belfast, meaning there was more Pb to retain in the soil. Therefore, if OM is not included in the Glasgow RF model, the validation statistics in Belfast and Leicester improve ($R^2$ increase of 0.03 and 0.06 for Belfast and Leicester respectively), although the internal validation statistics are considerably worse if OM is not included in the Glasgow model ($R^2$ decrease of 0.04). In contrast, other covariates such as building age and historic industry density, have a relatively consistent relationship with soil Pb, regardless of geographical location. Therefore, including building age and historic industry density in the RF model will likely improve its accuracy in other cities.

This trade-off between covariate inclusion or combinations, and the accuracy of the RF model in other cities, could have been experimented with in more detail. Machado et al. (2019) demonstrated how different training datasets and covariate combinations can influence RF accuracy in other regions, when predicting soil classes in Brazil. Similarly, Orynbaikzyzy et al. (2022) trained their RF model using samples from multiple regions, otherwise the model accuracy was too low. However, wide-scale testing of covariate inclusion on external validation statistics in this study would have essentially resulted in the model being trained by all the G-BASE/Tellus datasets. This was not the aim of this study. Instead this research has highlighted the importance of carefully selecting appropriate covariates to be included in a RF model and being aware of model transferability. Overall, the trade-off between a general-purpose model, against a model which has high local accuracy needs to be recognised, especially for urban soil which are complex to model.
10.3.3. Comparison of techniques

Arguably, the most important output of this thesis has been the maps of soil Pb, with different techniques used to generate maps with different emphases. This has involved mapping soil Pb concentrations, and also soil Pb risk, for different scales and via different techniques. These maps help us answer the third and fifth research questions i.e. can we map soil Pb concentrations and locate areas posing a greater risk using different modelling methods? (Section 1.2). These properties have been mapped using model-based geostatistics i.e. OK and E-BLUP (Chapters 5 and 6); and machine learning i.e. RF/QRF (Chapter 8), with different scales and levels of accuracy found for each technique (Figure 10.1).

Geostatistics and machine learning use the same datasets but in very different ways. For example, geostatistics relies on nearby measured values (i.e. soil Pb concentrations) to predict values at unsampled locations (Equations 5.8 and 5.9), with the inclusion of linear relationships with covariates in order to improve prediction accuracy (Equation 6.5). On the other hand, machine learning relies on the covariate values at each unknown location to estimate soil Pb concentrations or risk, with measured values used to train and validate the model (Figure 8.2).

In contrast to geostatistical methodologies, machine learning techniques accommodate complex nonlinear relationships between measured values and covariates. Therefore, geostatistical techniques tend to lead to smoother predictions, as they are a weighted average of nearby measurements (Figure 10.1). The simplicity of the linear relationships means that the causes of variation can be more easily interpreted, and the explicit modelling of spatial dependence means that hypotheses regarding drivers of variation can be tested. However, machine learning models are potentially transferable as they can be applied to other study regions where only covariate datasets can be accessed (Section 10.3.2). They are also considerably more flexible in terms of the types of relationships that can be accommodated e.g. non-linear.
### SCALE

**City:** Greater Glasgow (G-BASE)  
**Neighbourhood:** Paisley and Bishopbriggs

#### Geostatistics
- Ordinary kriging (log Pb)
- Transparent to show uncertainty (mg kg⁻¹)
- Classified by C45L exceeded to show risk

#### Machine Learning
- Random forests (RF; log Pb)
- QRF to show uncertainty (C45L classes)
- Soil Pb risk - probability high soil Pb plus hotspots

**Figure 10.1:** Gallery of main soil Pb map results, sorted by scale (city versus neighbourhood datasets) and mapping technique (geostatistics versus machine learning). With figure numbers provided for further reference.
Both geostatistics and machine learning techniques use covariates to predict soil Pb concentrations. However, the relative influence or importance of covariates on soil Pb concentrations varies slightly between the methods (Figure 10.2). Although the direct influence of each covariate is difficult to compare between methods, the relative importance of the covariates can be ranked from most to least influential, depending on their Wald value or predictor importance estimate (Figure 10.2). Most covariates are not influenced by the techniques used, for example OM (log) ranks the most influential covariate regardless of technique or scale. However, the importance of other covariates is influenced by the method used. For example, at the neighbourhood scale, building age ranks 3\textsuperscript{rd} most important when using geostatistics, but has the lowest influence (ranked 5\textsuperscript{th}) in the Paisley RF model. Overall, the technique used can have a minor impact on the importance of some covariates when modelling soil Pb. However, scale and/or location play a bigger role, as discussed previously (Sections 10.3.1 and 10.3.2).

In summary, geostatistical models are more interpretable but machine learning models permit more complex relationships. Therefore, machine learning models can be more accurate providing appropriate covariates are selected (see Section 10.3.2). Scale also plays an important role in both techniques, as discussed in Section 10.3.1. It would have been interesting to compare more mapping techniques and outcomes from different machine learning methods, which is discussed in more detail in Section 10.4.1. Overall, which technique should be used to generate a map of soil Pb concentrations or risk depends on the mapping scale, available datasets, and end goals of the map. There are also different ways to communicate the resulting outcomes using both techniques. This is discussed in the next section.
**SCALE**

**City:** Greater Glasgow (G-BASE)  
**Neighbourhood:** Paisley and Bishopbriggs

**Table 6.10:** Summary of ranked Wald and p values for each covariate. *R* = Wald rank, H.I. = historic industry. *">95%" = >99% significance.

<table>
<thead>
<tr>
<th>Wald</th>
<th>R</th>
<th>p</th>
<th>Covariate</th>
</tr>
</thead>
<tbody>
<tr>
<td>507</td>
<td>1</td>
<td>&lt;0.01**</td>
<td>OM (log)</td>
</tr>
<tr>
<td>42</td>
<td>2</td>
<td>&lt;0.01**</td>
<td>Building age</td>
</tr>
<tr>
<td>21</td>
<td>3</td>
<td>&lt;0.01**</td>
<td>H.I. density (log)</td>
</tr>
<tr>
<td>13</td>
<td>4</td>
<td>0.0012**</td>
<td>Texture</td>
</tr>
<tr>
<td>12</td>
<td>5</td>
<td>0.018*</td>
<td>H.I. age</td>
</tr>
<tr>
<td>11</td>
<td>6</td>
<td>0.055</td>
<td>Land use</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>0.032*</td>
<td>Road type</td>
</tr>
<tr>
<td>4</td>
<td>8</td>
<td>0.057</td>
<td>H.I. type</td>
</tr>
<tr>
<td>2</td>
<td>9</td>
<td>0.131</td>
<td>Building Dist. (log)</td>
</tr>
<tr>
<td>1</td>
<td>10</td>
<td>0.27</td>
<td>pH</td>
</tr>
<tr>
<td>0.02</td>
<td>11</td>
<td>0.88</td>
<td>H.I. Dist. (log)</td>
</tr>
<tr>
<td>0.01</td>
<td>12</td>
<td>0.93</td>
<td>Road Dist. (log)</td>
</tr>
</tbody>
</table>

**Table 6.12:** The relationship between soil Pb and each covariate in Paisley and Bishopbriggs.

<table>
<thead>
<tr>
<th>Area</th>
<th>Covar.</th>
<th>Wald</th>
<th>R</th>
<th>p</th>
<th>Wald</th>
<th>R</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paisley</td>
<td>OM (log)</td>
<td>57</td>
<td>1</td>
<td>&lt;0.01**</td>
<td>92</td>
<td>1</td>
<td>&lt;0.01**</td>
</tr>
<tr>
<td>Land use</td>
<td>24</td>
<td>2</td>
<td>&lt;0.01**</td>
<td>19</td>
<td>2</td>
<td>&lt;0.01**</td>
<td></td>
</tr>
<tr>
<td>Road dist.</td>
<td>8</td>
<td>5</td>
<td>&lt;0.01**</td>
<td>15</td>
<td>3</td>
<td>&lt;0.01**</td>
<td></td>
</tr>
<tr>
<td>Build dist.</td>
<td>10</td>
<td>4</td>
<td>&lt;0.01**</td>
<td>6</td>
<td>4</td>
<td>0.016*</td>
<td></td>
</tr>
<tr>
<td>Build age</td>
<td>14</td>
<td>3</td>
<td>&lt;0.01**</td>
<td>1</td>
<td>7</td>
<td>0.93</td>
<td></td>
</tr>
<tr>
<td>H.I. dist.</td>
<td>0</td>
<td>7</td>
<td>0.97</td>
<td>3</td>
<td>5</td>
<td>0.073</td>
<td></td>
</tr>
<tr>
<td>H.I. type</td>
<td>1</td>
<td>6</td>
<td>0.71</td>
<td>3</td>
<td>6</td>
<td>0.57</td>
<td></td>
</tr>
</tbody>
</table>

**TECHNIQUE**

**Machine Learning (RF)**

**Figure 10.2:** Summary of covariate importance separated by scale (city versus neighbourhood datasets) and mapping technique (geostatistics versus machine learning). Wald values are generally ranked from highest to lowest in the tables, and predictors are sorted by Wald values (highest at the top or on the left), with predictor importance rank shown in orange.
10.3.4. Communication of results and end uses

The final research question aims to determine the best methods for communicating the results from soil Pb modelling, both via kriging and predictive modelling. This is vital to consider as it will determine how useful the results are. However, the end user of the map often dictates which communication method is the most effective.

Figure 10.1 provides a summary of the different types of maps generated from the modelling methods. Both types of mapping e.g. OK and QRF, initially produce logarithmic values of soil Pb (e.g. Figures 5.6A and 8.4). These logarithmic maps display the spatial patterns in soil Pb values, so identify soil Pb hotspots and coldspots, but they are not easily interpreted. Therefore, the values were classified by which of several land use C4SL they exceeded (e.g. Figures 5.8A and 8.5). However, for RF predictions in Greater Glasgow (Figure 8.5) only one to three C4SLs are likely to be exceeded, depending on which percentile is selected. This homogeneity is partly due to the central tendency issue with RF predictions.

The maps in Figure 10.1 also represent uncertainty in different ways. For example, Figure 5.6 shows the kriging prediction alongside kriging error. Similarly, in Figures 8.5, 8.14, and 8.15 the upper and lower confidence intervals maps are presented beside the median value prediction map to show uncertainty in the predicted values. Side-by-side or static comparison allows the user to compare how the predicted values, and the uncertainties in those predictions, vary from cell to cell, with ‘data’ on the left and ‘uncertainty’ on the right of the figure (MacEachren et al., 2005). However, the side-by-side display can take up valuable space on the page and divert the reader’s attention from the main message. Side-by-side display was favoured by decision makers in urban planning issues, according to a web-survey conducted by Aerts et al. (2003).

An alternative method for conveying uncertainty is to use transparency (e.g. Figures 5.7 and 5.8; MacEachren et al., 2005). It is more subtle than side-by-side maps; therefore, it can be difficult to interpret the spatial patterns in uncertainty values. However, soil Pb’s spatial pattern and its associated uncertainty can be conveyed in one map, which may be preferred by decision makers. Bisantz et al. (2011) found transparency was an important tool when communicating the uncertainty of missile strikes to decision makers.
The soil Pb maps generated from this study (Figure 10.1) have various different potential end users. This is discussed in Section 8.4.3., where it is suggested that the probability maps (e.g. Figures 8.8 and 8.18) could be incorporated into a desk-based study before new developments, whereas the C4SL classified maps (e.g. Figures 8.5, and 8.14, and 8.15) could help inform a local authority of the most appropriate future land uses. It is important to emphasise that the machine learning maps should only be used as a starting point from which to carry out further investigation, e.g. to decide if or where soil samples should be collected.

One potential end user of the soil Pb machine learning maps are landowners and building developers. The UK government recommends undertaking a land contamination risk assessment (LCRM) before changing the current use of a site (McMahon et al., 2001). The first step of a LCRM involves a preliminary environmental risk assessment where a desk-based study is often used to screen for any potentially unacceptable risks likely to be present on a site (Crown copyright, 2021). This may include examining historical data or maps to determine the previous uses of the site and to identify possible contaminants present (McMahon et al., 2001). This aims to build a conceptual site model where potential source-pathway-receptors linkages are identified (McMahon et al., 2001).

The use of historical maps to identify risk can be subjective and dependent on available data. Therefore, machine learning modelling, in particular QRF, offers an alternative to assess the likelihood of a site containing a potential contaminant. This was demonstrated by Wragg and Cave (2021), who used BGS data from Stoke-on-Trent to spatially predict total and bioaccessible soil Pb via RF. They intend that their results could be directly input into contaminated land risk assessment models and could be applied to other urban centres. Furthermore, they state that this will support future urban land-use planning and land repurposing.

Overall, there are several different ways of visualising the same modelling results. For example, using logarithmic values, back-transformed concentrations, or comparison with C4SLs. Furthermore, the uncertainty in these predictions can be communicated using techniques like side-by-side maps or transparency. The end user often dictates the most useful communication method. These end users could include local authorities planning future land use or pinpointing current areas of concern, or local developers undertaking a desk-based survey.
10.4. Further work

Several avenues of further research have been highlighted by this thesis. This includes examining other modelling techniques, alternative covariate selection, modelling of additional elements in the soil, and modelling exposure routes.

10.4.1. Other modelling techniques

As mentioned in Section 10.3.3, it would be interesting to compare the RF accuracy results with other types of modelling methods e.g. machine learning methods such as artificial neural networks, support vector machines, logistic regression, Bayesian networks, and classification trees; and more traditional modelling methods such as generalised linear mixed or additive models. However, this type of comparison has been undertaken in other studies (e.g. Pranckevičius and Marcinkevičius, 2017; Han et al., 2018; Khoirunissa et al., 2021; Sun et al., 2021), who generally found that RF models produced the highest accuracy. Nevertheless, these studies did not use datasets from the field of soil contamination.

Obeng-Gyasi et al. (2021) performed a comparative study of machine learning techniques to predict soil Pb concentrations. They did not use RF, but instead used other types of machine learning tools, i.e. logistic regression and Bayesian networks, to predict where soil Pb concentrations were likely to exceed 100 mg kg\(^{-1}\) in front gardens in Greensboro, North Carolina. Their AUC ranged from 0.74 to 0.83, indicating a high degree of accuracy. This demonstrates the capacity for other machine learning techniques at predicting soil Pb concentrations.

10.4.2. Other covariates

Another potential possibility for future research could be using more remotely sensed data in RF models to predict soil Pb. Remotely sensed datasets are often easier to obtain than anthropogenic covariates, so they have been included in many other RF models (see Table 8.1 e.g. Kirkwood et al., 2006; Tan et al., 2019; Omondi and Boitt 2020). However, these studies were all undertaken in rural regions, where there will be less overprinting of Pb from multiple sources. Therefore, remotely sensed covariates have not been included in RF models predicting soil Pb in British urban environments, to the authors current knowledge. It would also be interesting to explore if remotely sensed variables correlated with soil Pb concentrations at different scales in Glasgow, and how much of the correlation was due to the influence of OM.
Recent machine learning research has also included socio-economic covariates in their predictive soil Pb models. For example, Obeng-Gyasi et al., (2021) found that soil Pb risk (i.e. the likelihood of soil containing > 100 mg kg\(^{-1}\) of Pb) was double in 100% black neighbourhoods, compared with 100% white parts of Greensboro, USA. Similarly, O’Shea et al. (2021) found strong correlations between soil Pb concentrations, the percentage of children with elevated blood Pb levels, the percentage of minority populations, as well as the percentage of children in poverty, in Philadelphia, USA.

Although the relationship between soil Pb and blood lead has been extensively studied in the USA (e.g. Mielke et al., 2022), there has not been as much research into the link between soil Pb and socio-economic factors in the UK. This could be explored using the Scottish Index of Multiple Deprivation (SIMD) which combines multiple deprivation factors e.g. housing quality and employment rates, into one measure (Gov.scot, 2020). Therefore, areas of a city with a higher SIMD tend to have older and poorer quality housing, which may link to soil Pb concentrations. For instance, Parkhead in the Eastend ranks in the top 5% most deprived areas in Scotland (Figure 10.3), and this area is predicted to be a soil Pb hotspot (Figure 8.8). It would be interesting to formally test if areas with higher soil Pb concentrations have a higher SIMD in Glasgow, and if so, what is driving this relationship.

![Figure 10.3: A screenshot from the SIMD 2020 viewer focused on Glasgow, with Parkhead West and Barrowfield data zone shown on the right. Red data zones = higher SIMD rankings so are more deprived. From SIMD.scot (2020).](image-url)
10.4.3. Other elements

Concentrations of other soil elements were available in the Glasgow datasets; as the BGS measured the concentrations of approximately 50 soil substances in the G-BASE samples (Fordyce et al., 2017). Correspondingly, I used a multi-element standard when running ICP-OES analysis on the samples from Paisley and Bishopbriggs (see Section 3.2.3.1). This determined the concentrations of 22 additional soil elements including Cd and Cr.

The methodology developed in this thesis could also be applied to other potentially toxic elements (PTE) in the soil e.g. training a RF model for As, Cd, Cr, and Hg. This would provide a more complete picture of potentially toxic elements in the soil, which would be useful when conducting a desk-based environmental assessment. However, careful consideration would have to be given to which covariates to include in a RF model predicting multiple soil elements, as it is unlikely that one size will fit all.

Several other studies have used RF models to predict multiple elements in the soil. For example, Radocaj et al., (2020) used a RF model, along with other methods, to predict several heavy metals (Cd, Cr, Cu, Ni, and Zn) in soils throughout Mainland Croatia. Similarly, Xiao et al. (2020) used a RF model to assess PTEs (Cd, Hg, As, Pb and Cr) distribution and controls in Xiangzhou, China. However, these studies focus on rural areas and did not use QRF to quantify uncertainty.

CoDA methods would be highly applicable for analysing multiple elements due to the constant sum issue (see Section 2.6.). This most commonly involves calculating the log-ratio between elements, with multiple log-ratio methods suitable for geochemical data (McKinley et al. 2016). These include; pairwise log-ratio transformation (pwlr), addictive log-ratio transformation (alr), centred log-ratio transformation (clr), and isometric log-ratio transformation with dendrograms (McKinley et al. 2016).

One potentially interesting use of CoDA would be calculating and mapping the ratio between Pb and Mo in the soil. Molybdenum is heavily used to lubricate brake pads (Folkeson, 2005); therefore, soils containing high Mo concentrations may indicate that Pb in these soils predominantly originated from car/petrol sources.
10.4.4. Other exposure routes and risk modelling

As discussed in Section 2.2., there are many potential exposure pathways of humans to Pb e.g. ingestion or inhalation of soil, food, water, and air/dust (EA, 2009). Each of these exposure mediums have been modelled individually, for example Pb in food (Brown et al., 2016; Entwistle et al., 2019), water (Chojnacki et al., 2017; Fasaee et al., 2021), and air (Sethi and Mittal, 2019). However, more recent studies have combined multiple Pb mediums and routes in one model, which is a more powerful tool for combating Pb health issues. For instance, Lobo et al. (2021) used machine learning methods (including RF) to predict children’s blood Pb levels using socioeconomic, housing, and water quality covariates.

Expanding this study to consider additional exposure sources would be challenging as several extra datasets would be required e.g. Pb measurements for food, water, air, dust, and even blood levels. These datasets would be difficult to obtain.

Alternatively, the risk from soil Pb could have been included in the RF model. For example, Wragg and Cave (2021) predicted bioaccessible Pb, in addition to total Pb, using RF models in Stoke-on-Trent. As discussed in Section 2.2. not all ingested on inhaled Pb will enter the human bloodstream i.e. is bioaccessible. Therefore, predicting bioaccessible, instead of total Pb, will give a better indication of the potential health risk from soil Pb. However, total and bioaccessible Pb are closely linked, and bioaccessible Pb can be modelled from total Pb concentrations (e.g. Appleton et al., 2012).

Other studies have used the results from machine learning to calculate the exposure risk from soil Pb (e.g. Pelfrène et al., 2015; Xiao et al., 2020; Chen et al., 2022). This may involve calculating the hazard quotient (HQ) from a specific element (USPEA, 1989), which is achieved by calculating the ratio between the estimated daily intake (EDI) and the tolerable daily intake (TDI, provided by USEPA). If the HQ is > 1 then there could be adverse health effects from Pb. The EDI is calculated by:

\[
EDI = \frac{\text{Soil Pb} \times IR \times EF \times ED \times \text{Bioaccessible Pb} \times 10^{-6}}{ABW \times AT},
\]

where \(IR\) is the soil and dust ingestion rate (mg day \(^{-1}\)) and is commonly obtained from USEPA guideline values. The averaging time \((AT)\) is the average intake over
the exposure duration (USEPA, 1989). $ED$ and $EF$ represent the exposure duration (years) and frequency (days/year) respectively and depend on the median age of the child (USEPA, 2009). Furthermore, average body weight ($ABW_{kg}$) is a denominator in the calculation of EDI, so exposure risk is highly dependent on the child’s age. Overall, the HQ tends to be greater for younger children, and calculated HQ for different age groups can combined to assess the exposure risk throughout childhood (Reis et al., 2013).

The HQ could be calculated from the RF predicted soil Pb concentrations (and related bioaccessible Pb); however, the error margin is likely to be high. Other exposure risk calculations are available e.g. the contaminated land exposure assessment (CLEA), but also require detailed and accurate datasets.

10.5. Conclusions

This thesis has met its original aim to develop a suitable predictive model for soil Pb concentrations in British urban environments. Furthermore, this model can be used to locate the areas of a city most at risk from high soil Pb concentrations. This is an important starting point for future studies and remediation projects, which will hopefully mitigate the potential health impacts from high soil Pb concentrations.

This project has developed new and transferable techniques for modelling soil Pb concentrations. For instance, QRF were used to predict the risk and uncertainty in soil Pb values, especially around soil Pb hotspots. In particular, higher percentiles (e.g. the 95th percentile) more accurately predicted higher soil Pb concentrations (i.e. $\geq 200 \text{ mg kg}^{-1}$), based on internal cross-validation statistics. The QRF percentiles also generated a range of likely soil Pb concentrations (i.e. upper and lower confidence intervals). Lastly, the percentile at which a high soil Pb threshold (e.g. 200 mg kg$^{-1}$) is exceeded, was used to calculate the probability that a location will contain high soil Pb values.

This research benefits from analysing soil Pb concentrations at different scales in the same city. Therefore, the importance of modelling scale was considered, in respect to both the scale of spatial variation and the measurement scale. Spatial variation in soil Pb concentrations can be witnessed at every scale in Glasgow, with samples showing spatial dependence from 50 m to 10.5 km. This agrees with other research where heterogeneity in soil Pb values has been observed from the ultra-fine scale to the continental scale. This range in spatial variation is largely controlled.
by the measurement scale, which can influence the mapping resolution and the relationship between soil Pb concentrations and covariates. Therefore, in this study different covariates were used at varying scales, with covariate selection based on initial geostatistical analysis. Overall, the sampling and mapping scales are fundamental parts of a study modelling soil Pb concentrations, so should be carefully considered before analysis.

This study is also unique as it tests the transferability of the Glasgow QRF models. Machine learning model’s main benefit is that they are transferable, i.e. they should be able to predict soil Pb concentrations, and risk, in other geographical locations. However, this is rarely examined. Model transferability was firstly found to depend on the number of covariates included in the model; with more covariates increasing model accuracy in the training city but reducing model accuracy in other geographical locations. However, model transferability can also be controlled by the consistency of the relationship between soil Pb and the covariates in other places; with some covariates showing a similar relationship with soil Pb regardless of location, and other covariates exhibiting larger variations. On the whole, future soil Pb machine learning models should be wary when selecting covariates in the training dataset.

Several different mapping methods were used to generate a map of soil Pb concentrations and risk. For instance, OK, E-BLUP, RF and QRF all produced maps of soil Pb using different techniques. It is difficult to directly compare these outcomes, and the choice of method depends on available datasets, mapping scale, and end goals of the map.

Lastly, consideration was given to how the results from RF and QRF predictions should be displayed depending on the end user. For example, predicted soil Pb concentrations were compared with C4SL to help inform a local authority or homeowner on the most suitable future land uses for an area of land. Similarly, predicted soil Pb hotspots were emphasised using contour lines to indicate the areas of a city or neighbourhood most likely to be at risk from high soil Pb concentrations. Uncertainty was also communicated using different methods e.g. side-by-side maps and adjusting the symbol transparency. Overall, it is hoped that mapping based on machine learning predictions will be incorporated in some way into desk-based environmental risk assessment studies, that are undertaken before
changing the use of a site. This will assist with future land-use planning and repurposing.

There are several different avenues of research that could be explored based on this thesis. For instance, the accuracy of QRF predictions of soil Pb could be compared with other modelling techniques e.g. ANN and SVM. The QRF predictions may also be improved by including other covariates, perhaps from remote sensing datasets or using more socio-economic variables e.g. the SIMD. The methodology developed in this thesis could also be applied to other potentially toxic elements in the soil e.g. Cd and Cr. Similarly, other exposure pathways of Pb could be modelled e.g. via ingestion/inhalation of food, water and the air, and the total exposure risk from Pb could be predicted and visualised.
References - Chapter 10


