Investigating Lead Contamination in the Glasgow City Region.

Subasish Behera

Contents

1 Introduction **2**

1.1 Project Background 2

1.2 Aims and Objectives 2

1.3 Description of the data 2

2 Description of Statistical Methods **3**

2.1 Ordinary Least Squares Linear Model 3

2.2 Spatial Model: accommodating spatial autocorrelation 4

3 Analysis **9**

3.1 Data preparation 9

3.2 Spatial Trends and Clusters 9

3.3 Effects of Co-variates on Spatial Pattern 14

3.4 Outlier Analysis 16

4 Research Conclusion and Discussions **16**

5 References **16**

1 Introduction

1.1 Project Background

The research project undertakes an investigation of top-soil concentration of the heavy-metal Lead (Pb) in the Glasgow City Region. Small amounts of Lead can be found naturally in the soil, but concentrations are enhanced by man-made activities. Research suggests that lead contaminated soil and dust are the most important factors of high blood lead level(Lanphear, B. P. et al., 1997). Adverse health problems from high blood lead level has been well documented through out the world. Chronic exposure to low-level lead has been linked to several developmental problems, especially for pre-school children(Needleman H. L., et al., 1990). There is still much on-going research to solidify the relationship between lead amount found in soil and amount that is absorbed by humans through exposure.

1.2 Aims and Objectives

The primary objective of this report is to investigate the processes that lead to accumulation of Pb in the top-soil of the Glasgow City Region. This broad objective can be divided into four sub-objectives that will provide more insight into the underlying factors:

1. To identify the spatial pattern of Pb contamination in the Glasgow area.

2. To identify clusters of contamination, if there exists any.

3. To quantify the effects of co-variates on the Pb concentration.

The objectives render the research problem as firstly a spatial problem where the concentration values are mapped across the study region. This is followed by a regression problem which may or may not involve adjusting for a spatial component depending on the strength of correlation between response variable. The completion of the objectives has social-importance as it will help the officials investigate highly contaminated areas and use the predictions to identify potentially hazardous areas. This is aimed at sustainable development and planning for the Glasgow City Region.

1.3 Description of the data

The study area includes each of the eight local councils that constitute the Glasgow City region, i.e. Glasgow City council, Invercylde council, Renfrewshire council, East Renfrewshire council, East Dunbartonshire council, West Dunbartonshire council, North Lanarkshire council, and South Lanarkshire council. The data used in this project is a subset derived from a wide range of datasets collected by the British Geological Survey (BGS) between 2001 and 2011 for the Geo-chemical Baseline Survey of the Environment (G-BASE) project. This served as the foundation for the multi-disciplinary project titled 'Clyde Basin Urban Super Project' (CUSP) which was carried out in an attempt to provide sustainable planning and development in Scotland's major urban areas.

The soil survey in the Clyde Basin for the CUSP was conducted in two parts.The first part included sample collection from the urban and sub-urban area of Glasgow city in 2001-2002. The second part included sample collection from the remaining urban cities in the study region and also the rural areas.

The Co-ordinate Reference System used for this data is British National Grid (EPSG: 27700) which presents the co-ordinates in Easting and Northing. The response variable under study is the top-soil Pb concentration in the Glasgow City Region, which is measured in parts per million (ppm). Along with the concentration value, other co-variates relating to the location of sample were also collected. The co-variates involved and their short description is given in Table 1.

Table 1: Description of variables in the dataset

Variable Name	Description
X	Eastings (in meters)
Y	Northings (in meters)
Pb	Lead concentration (in ppm)
Elevation	Sample height above or below the mean sea level (in meters)
Slope	Change in elevation over a certain distance
Aspect	Orientation of maximum slope measured clockwise
Plan.Curvature	Affects convergence/divergence of flow across the surface
Profile.Curvature	Affects acceleration/deceleration of flow across the surface
TWI	tendency of area to accumulate water
MRVBF	characterizes flat bottom areas
MRRTF	characterizes high flat areas
Population	population density per km^2
Landuse	categorical with levels: 1, 3, 4, 7, 9, 10, 11, 12, 13, 14, 19, 20, 21,

All the inherent variables are continuous except `Landuse`*. The context of the levels of* `Landuse` *are not provided. The number of categories matches with that specified in National Land Use Database(NLUD) project. However the mapping of codes to its meaning is unknown and there is a mismatch between the total categories available here and those used in the original report. The link to the original report published by the BGS and the National Land Use Database project's guideline is provided in the section* 5.

2 Description of Statistical Methods

2.1 Ordinary Least Squares Linear Model

Linear regression is referred to as a parametric approach for modelling a scalar response variable (also referred to as dependent variable) using one or more explanatory variables (also called independent variables). This statistical model models the linear relationship between the response and explanatory variables and takes the parametric form:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + ... + \beta_p x_{pi} + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2) \quad i = 1, ..., n$$

The above equation is often expressed in a very convenient vector-matrix form as follows:

$$Y = X\beta \ + \ \epsilon, \quad where\ E(\epsilon) = \vec{0}\ and\ Var(\epsilon) = \Sigma$$

also,

$$\sum = \begin{bmatrix} \sigma^2 & 0 & 0 & ... & 0 \\ 0 & \sigma^2 & 0 & ... & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & ... & \sigma^2 \end{bmatrix}$$

The Σ is termed as the associated variance-covariance matrix of the error terms.

Now, it is also mathematically equivalent to notate the above information as follows:

$$Y \sim \mathcal{N}(X\beta, \Sigma)$$

The beta parameters are estimated by the method of least-squares, which minimizes the sum of squared errors given by:

$$S(\beta) = \sum_i (y_i \ - \ x_i^T \beta)^2 \ = \ (Y - X\beta)^T (Y - X\beta)$$

The minimization of the above equation gives the estimated beta parameters, which in vector-matrix notation is given by:

$$\hat{\beta} = (X^T X)^{-1} (X^T Y)$$

By extension, the estimated errors (residuals) is calculated as $\hat{\epsilon} = (Y - X\hat{\beta})$

It is best practice to include an interval of estimates which provides a range of plausible values with some confidence, rather than a single estimate. The interval, called the confidence interval is calculated for the beta parameters as follows:

$$\hat{\beta}_i \pm \ t(n-p, \frac{\alpha}{2}) \sqrt{(X^T X)^{-1}_{ii}}$$

where, t(.) refers to the Student's t-distribution with $n - p$ degrees of freedom, α refers to the significance level (often, though not necessarily set to 5%), S.E(β_i) refers to the standard error of the beta parameters.

There are a certain number of assumptions associated with this modelling approach, which are stated as:

- The mean of the error terms ϵ_i is zero.
- The variance of the error terms is constant for all values of independent variables.
- The errors are independent for all observations.
- The errors have a normal distribution.

The assumption of independent errors gives a certain structure to the variance-covariance of the population under study. The off-diagonal elements of the variance-covariance matrix(given by Σ above) is 0 indicating independence of the errors. This assumption of independent errors is certainly not true for situations where the data under study has some temporal or spatial correlation. The correlation of the values of response variable indicates that when a statistical model is employed to model these data, the residuals generated will have some degree of correlation and the off-diagonal elements in the variance-covariance matrix will not be zero anymore, but rather will have a value dependent on the inherent correlation structure of the data (which is often unknown and estimated).

2.2 Spatial Model: accommodating spatial autocorrelation

When the data used in a study has an inherent spatial correlation, the assumption of independent errors is violated and thus, is needed to be addressed. The way to incorporate the spatial correlation is discussed.

The first Law of Geography, as stated by Waldo Tobler goes as follows:

"*Everything is related to everything else, but near things are more related than distant things.*"

This adage serves as the foundation upon which spatial modelling resides. So, the degree to which the two values will be correlated in a spatial setting will depend on how far the two values reside in space i.e., the distance. There are a number of distance metrics available that allows a practitioner to quantify how far the values reside, and again, depending on the type of data being modeled, one is preferred over another.

2.2.1 Geostatistical Process:

The spatial process appropriate and therefore used for the data in hand is called the 'geostatistical process', which is defined as:

$$Y(\mathbf{s}) : \mathbf{s} \in D$$

where D serves as a subset of the 2-dimensional space $\mathbb{R}^2$, i.e., the study region. The locations $\mathbf{s} = (s_1, s_2)$ varies continuously over the region D, but in practice only a sample of locations are selected and rest of the locations are generally needed to be predicted with some approximation of spatial smoothness.

The covariance of values at two locations is given by:

$$C(\mathbf{s}, \mathbf{t}) = Cov[Y(\mathbf{s}), Y(\mathbf{t})] = E(Y(\mathbf{s})Y(\mathbf{t})) - \mu(\mathbf{s})\mu(\mathbf{t})$$

where $\mu(\mathbf{.}) = E(Y(\mathbf{s}))$ i.e., the expected value of the process at that location.

2.2.2 Process Model:

The continuous response variables is assumed to follow a gaussian process. So, the vector of response values $[Y(s_1), Y(s_2) ... \ Y(s_n)]$ at n locations is denoted in a vector-matrix form as:

$$Y \sim \mathcal{N}(\mu, \Sigma)$$

for a mean vector μ and variance-covariance matrix Σ. The structure of the variance-covariance matrix is different from the one used in Ordinary Least Squares model. Using some generality, the structure of the variance-covariance matrix is given by:

$$\sum = \begin{bmatrix} \sigma^2 & f(d_{12}) & f(d_{13}) & ... & f(d_{1n}) \\ f(d_{21}) & \sigma^2 & f(d_{23}) & ... & f(d_{2n}) \\ \vdots & \vdots & \vdots & & \vdots \\ f(d_{n1}) & f(d_{n2}) & f(d_{n3}) & ... & \sigma^2 \end{bmatrix}$$

where d_{ij} represents the distance between i^{th} and j^{th} sample location and f is a function that maps that distance to some function that quantifies the correlation e.g. and exponential function. This function f incorporates all the necessary information about the correlation structure assumed to be present in the data. This variance-covariance structure allows a practitioner to include the correlation information between the values based on a distance metric and hence, the First Law of Geography is presented here in a quantified manner.

Certain assumptions are made about the geostatistical process before modelling the data. These include the assumption of weak-stationarity and isotropy.

The process is weakly-stationary if:

- it has constant mean throughout the study region i.e., the mean is not dependent on the location. Mathematically, $E(Y(\mathbf{s}) = \mu \ \forall \ \mathbf{s}$

- the covariance (and therefore, correlation) between two sample values is a finite constant and only only depends on the displacement vector between the two locations, which is mathematically stated as, $Cov[Y(\mathbf{s}), Y(\mathbf{t})] = C(\mathbf{h})$, where $\mathbf{h} = ||\mathbf{t} - \mathbf{s}||$.

A further assumption regarding the covariance is made which leads to 'isotropy'. The process is isotropic in nature if:

- the covariance between two sample values only depends on the euclidean distance between the two sample locations and is independent of the direction at which two sample locations are considered i.e., $C(\mathbf{h}) = C(\ ||\mathbf{h}|| = \sqrt{h_1^2 + h_2^2}\ )$.

The covariance structure of the Σ depends on certain parameters that are estimated from the data and thus, is written as $\Sigma(\theta)$, where $\theta = [\sigma^2\ \tau^2\ \phi]$.

- σ^2 is called partial sill and is defined as the amount of smooth variation in the data.
- τ^2 is called nugget effect and is defined as the amount of random variation present in the data.
- ϕ is called range parameter and is the defined as the distance at which the sample points become uncorrelated.

The most commonly used covariance models used to model the spatial autocorrelation are as follows:

$$\Sigma(\theta) = \sigma^2 \exp(-\mathbf{D}/\phi) + \tau^2\mathbf{I} \qquad (\textit{Exponential model})$$

$$\Sigma(\theta) = \sigma^2 \exp(-\mathbf{D}^2/\phi) + \tau^2\mathbf{I} \qquad (\textit{Gaussian model})$$

$$\Sigma(\theta) = \sigma^2[1 - \frac{3}{2}(\mathbf{D}/\phi)\ +\ \frac{1}{2}(\mathbf{D}/\phi)^3] + \tau^2\mathbf{I} \qquad (\textit{Spherical model})$$

The choice of covariance model is based on a model fitting criterion such as AIC, BIC, or by their predictive ability using cross-validation.

2.2.3 Detecting spatial autocorrelation

The correlation component in the spatial data is assessed using a Semi-Variogram. It is a plot of 'Semi-Variance' against distance which tells the practitioner whether or not a spatial model is adequate for the data. Semi-variance, in simple terms, is defined as a measure of dissimilarity between two sample values at a certain distance. Mathematically, semi-variance for a spatial process is stated as:

$$\gamma(s,t) = \frac{1}{2}\text{Var}[Y(s) - Y(t)]$$

As the semi-variogram is a population measure, an estimate is used in place, called binned empirical semi-variogram. The method to create a binned empirical semi-variogram is as follows:

- the statistical range of distances between all pairs of values respective to its co-ordinates is calculated and categorized into k intervals.
- For each interval, considering all values that lie within it, sum of squared difference between all pairs of points are calculated and is divided by $2|N(h_k)|$, where $|N(I_k)|$ is the total number of values or points within that interval, i.e.

$$\frac{1}{2|N(h_k)|}\Sigma[y(s) - y(t)]^2$$

- The above values are plotted against the midpoint of each interval.
- Along with the binned empirical semi-variogram, a Monte-Carlo envelope is calculated that simulates the upper and lower bound of semi-variance values expected at each distance under randomness.
- When this is calculated for each of the four direction, it is called a directional semi-variogram and helps to assess the isotropy assumption.

The way to interpret this estimate of semi-variogram is, if semi-variance values lie outside the estimated upper and lower bounds, then there is evidence of spatial auto-correlation in the data. For the directional semi-variogram, if the semi-variogram looks reasonably similar at each direction, isotropy assumption is validated. The binned empirical semi-variogram is only assessed at shorter distances (generally half the maximum distance) as the estimated bounds are unstable for larger distances.

There is a close relationship between the covariance structure and semi-variance as one determines the other. Thus, different covariance structures give different variogram models, which results to different shapes of binned empirical semi-variogram. Using the definitions, the relationship is expressed mathematically as:

$$\gamma(s,t) = \frac{1}{2}Var[Y(s) - Y(t)] \qquad = \frac{1}{2}[Var(Y(s) + Var(Y(t)) - 2Cov(s,t))]$$

Under the assumptions of isotropy and weak stationarity, the equation can be further reduced to:

$$\gamma(h) = C(0) - C(h)$$

2.2.4 Estimating Parameters

The spatial process

$$Y = [Y(s_1) \dots \ Y(s_n)] \sim \mathcal{N}(\mu(s), \Sigma(\theta))$$

has two classes of parameters that are estimated from the data. The mean function $\mu(s)$ is modeled via a linear approach and takes the parametric form $X\beta$ which is similar to the one seen in OLS model. The other class of parameters specific to the spatial model is the ones in the covariance structure i.e $\Sigma(\theta)$ as defined above.

The beta parameters are estimated in such a way that it incorporates the information of spatial correlation. The estimation can be conducted either by Maximum Likelihood Estimation or a Bayesian Approach, but the software used for the analysis of data in hand uses the Maximum Likelihood approach. The general form of the estimated beta parameters is as follows:

$$\hat{\beta}(\phi, \sigma^2, \tau^2) = (X^T V(\phi, \sigma^2, \tau^2) X)^{-1} X^T V(\phi, \sigma^2, \tau^2) Y$$

where, the expression for $V(.)$ varies for each covariance model, but it is always dependent on the three covariance parameters. The closed form solution is reminiscent of the OLS model solution, but differs from that as it includes the extra $V(.)$ term to include the spatial correlation information. The estimates of the (σ^2, τ^2, ϕ) are calculated from the data and plugged into the above formula to estimate the beta parameters.

The confidence intervals can be calculated after the model fitting to assess the uncertainty associated with the estimated parameter value:

$$\hat{\beta}_i(\hat{\phi}, \hat{\sigma}^2, \hat{\tau}^2) \pm \ t(n-p, \frac{\alpha}{2}) \sqrt{(X^T V(\hat{\phi}, \hat{\sigma}^2, \hat{\tau}^2) X)^{-1}_{ii}}$$

It is noteworthy that uncertainty intervals are only available for the mean model parameters i.e. beta parameters and no interval is generated for the covariance parameters.

The spatial model, by definition includes spatial autocorrelation and the estimates for the response values will be correlated depending on the strength of the correlation in the data as picked up the model, which implies that the residuals will also be correlated for the fitted values.Thus, the residuals need to be de-correlated and checked for independence, in order to assess whether spatial correlation is properly accounted by the model. This is done by decomposition of the variance-covariance matrix(e.g. by Singular Value Decomposition). The new residuals calculated are called 'innovations' and these will be used at the end to assess whether the model has properly accounted the spatial autocorrelation in the data.

2.2.5 Effects of Ignoring Spatial Correlation

The most important consequence of naively ignoring the inherent spatial auto-correlation that may be present in the data, is the underestimation of uncertainty intervals. The assumption of independence between the observations, when there is some correlation, reduces the coverage of the intervals.(Ferraciolli M. A., Bocca F. F., Rodrigues L. H. A.). This is because the model assumes there are more independent pieces of information than there actually is. Thus, this could mislead the practitioners to be more certain about the predictions, when more uncertainty is warranted.

2.2.6 Summary of the Model Fitting Steps

After the data cleaning and exploratory analysis phase, the analysis moves to model fitting. The steps involved in this phase can be summarized as follows:

- Fit an Ordinary Least Squares model to the data using a variable selection strategy assuming no spatial autocorrelation.
- Calculate the residuals, perform model diagnostics (except for the correlated residuals assumption) and re-iterate step 1 if required.
- After removing the linear trend due to co-variates via model fitting, plot the residuals against the co-ordinates to spot the inherent spatial trends and clusters.
- Using the residuals, also plot the binned empirical semi-variogram to look for evidence of spatial autocorrelation and interpret the plot accordingly.
- If evidence is found, use the spatial model to incorporate the autocorrelation. Using a model fitting criterion (e.g. AIC) choose the appropriate covariance model.
- After the spatial model is finalized and calculate the innovations. Use the innovations to create a binned empirical semi-variogram. Interpret the plot accordingly.

3 Analysis

3.1 Data preparation

The number of samples provided in the dataset is 2816. A skewness coefficient of 17.404 and the adjoining box-plot and density plot suggested a transformation of the variable before analysis. Log base-2 transformation was applied and the result of the transformation is presented in the graph below.

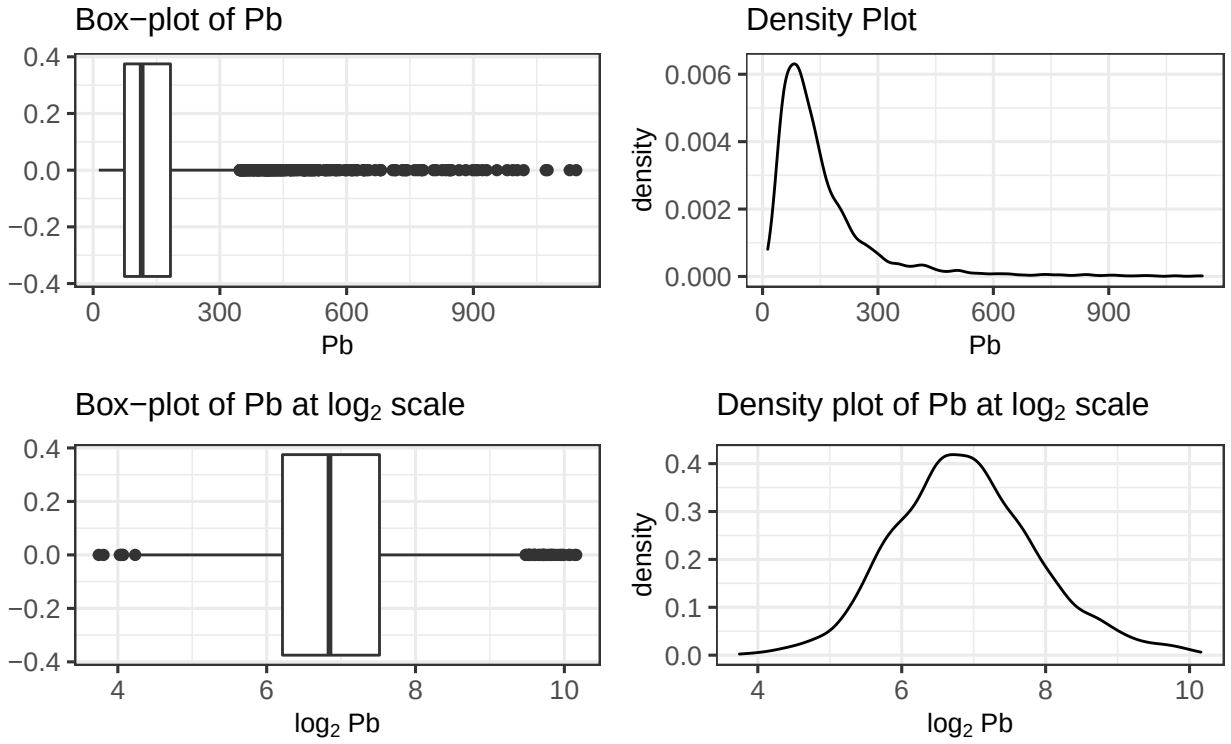
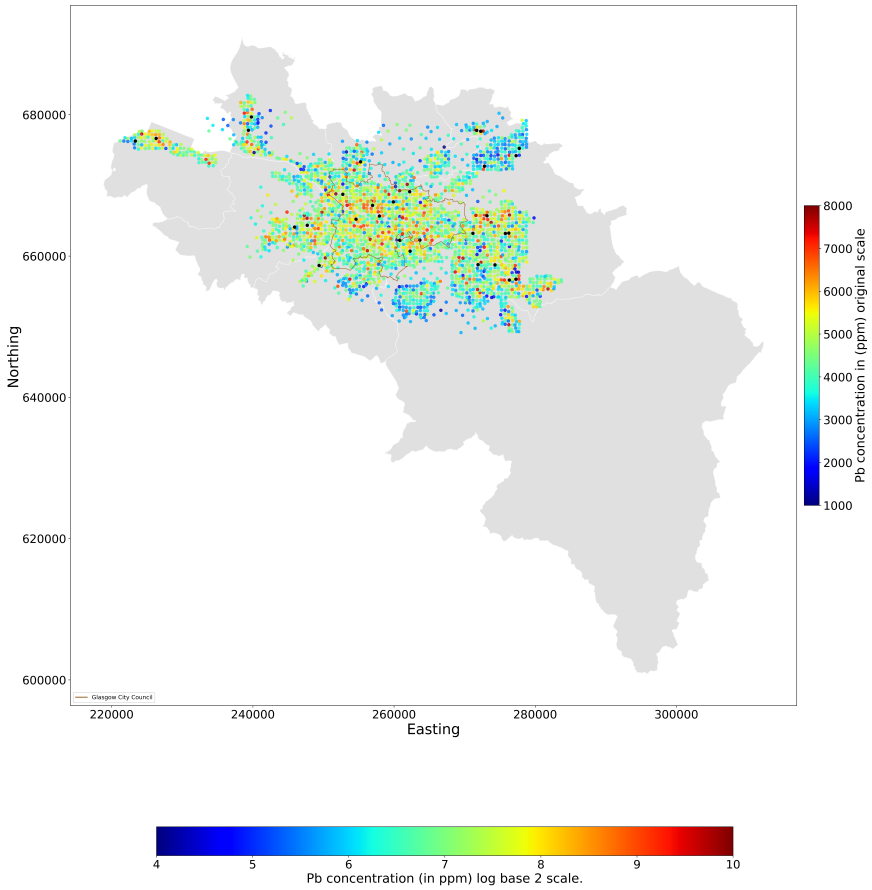


Figure 1: Effect of log_2 transformation on Pb variable

After the transformation, thirty outliers were spotted and were separated out to a different dataset to reduce noise for the statistical models. These outliers were analyzed differently after the analysis of non-outlier data points. 3-Standard Deviation rule was applied to locate an outlier. A total of 26 missing values were found in the response variable. Since the process being studied is roughly contiguous in nature, K-Nearest Neighbor using K = 5 was implemented to impute those missing values. Corresponding county information was collected from the shape file which will allow between and within county analysis. Based on evidence shown in plots 2 and 4, two new features were created and the rationale for which is explained in the section 3.2.

3.2 Spatial Trends and Clusters

The map of the Pb concentration with their corresponding sample location is provided below:



The data used in the plot is in log base-2 scale but another colorbar relative to original scale is also provided so as not to skew any interpretation. The locations of outliers are marked in black. There were no clusters of outliers which tells us that these unusual values were not from specific regions. Some features of the map are noteworthy. Evidently some clusters can be seen, specifically in the Glasgow City Region. A short term spatial correlation is also evident as sample locations near to each other have roughly similar values. No clear trend is apparent from this map.

The summary statistics for Pb concentration across county is given by the table 2. Some points to note are as follows:

- The sample size for each county is large, so the summary statistics and any relevant inferences made are valid.
- The median for Glasgow City Council is higher than any other counties.
- The variability for each county is roughly similar as shown by the standard deviation measure.

Table 2: Summary Statistics of Pb for each county.

county	Min	Q1	Mean	median	Q3	Max	Stdev	SampleSize
East Dunbartonshire	14.0	60.00	115.94	87.80	123.2	1019.0	115.75	202
East Renfrewshire	24.2	74.75	139.99	110.40	163.4	1075.6	124.22	155
Glasgow City	21.3	101.73	189.37	146.50	221.5	1143.3	145.46	698
Inverclyde	30.8	91.10	174.02	132.60	218.1	834.1	132.73	111
North Lanarkshire	16.8	68.02	149.17	109.55	179.0	1003.9	137.82	676
Renfrewshire	42.4	87.85	169.47	132.50	199.8	1071.4	134.79	295
South Lanarkshire	16.3	59.90	120.35	89.20	133.9	882.0	117.14	445
West Dunbartonshire	13.4	64.00	138.52	103.20	183.7	708.9	113.21	189

Because of the high skewness in the Pb concentrations, values at log scale are used for analysis hereafter unless specified otherwise. A 95% confidence interval along with the mean for each county is shown by the graph below.

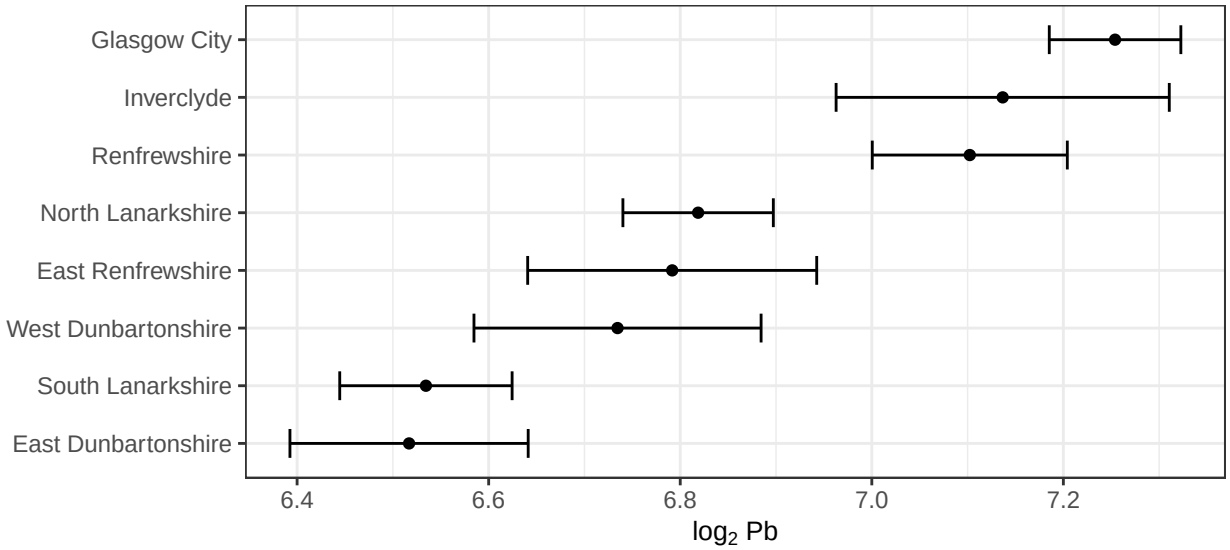


Figure 2: Mean Values with uncertainty interval across Counties.

Since log transformations are one-to-one transformations of the measure, the values can be directly converted to original scale by raising each of them to the power of 2. Direct inferences could be made on differences of means for each county based on whether or not there is overlap between the intervals. Glasgow City, Inverclyde, and Renfrewshire councils make a cluster in the high concentration side. This suggest feature engineering to create an indicator variable indicating whether the observation corresponds to these counties in order to allow the statistical models recognize the signals in the data. The variable is referred to as `High Concentration County` in the report wherever necessary. There are a total of 13 categories in `Landuse` variable used in the dataset. The frequency of the categories is depicted in the bar graph 3.

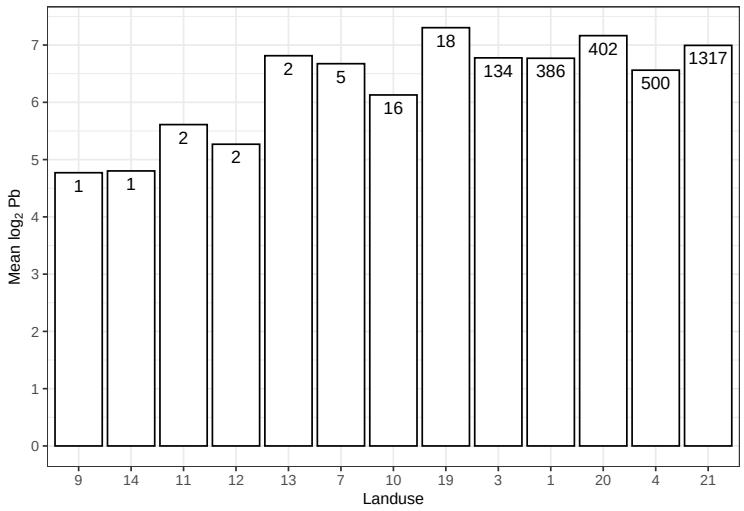


Figure 3: . Mean Pb concentration in log_2 scale with samples sizes labelled.

The average Pb concentration varies little in cases of high concentration categories, with the exception of `Landuse` category '19' , but the sample size is low in relation to statistical standards. The interaction of `Landuse` and `County` variables is well visualized by the graph 4 which gives a heatmap of concentration across `Landuse` and 'County' variable. Low sample size categories were removed before plotting.

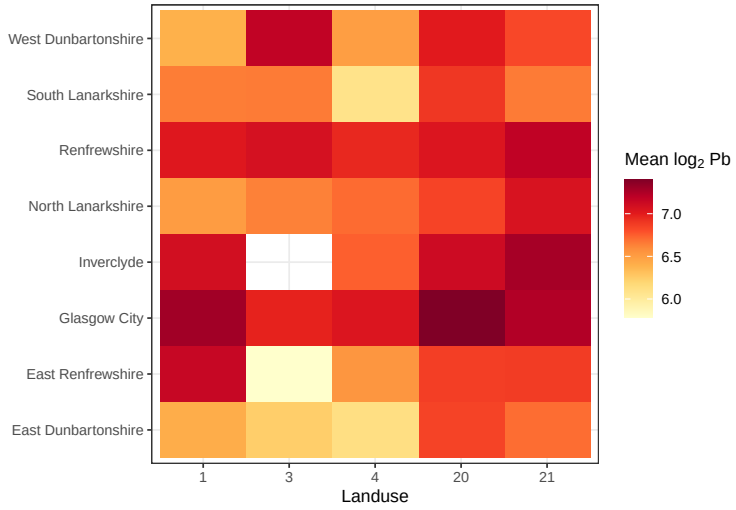


Figure 4: Heatmap of Pb concentation across Counties and Landuse

The categories 20 and 21 generally had high level of Pb concentration across all counties and, also had little differences in their summary statistics. Based on this fact, another feature was created indicating whether the `Landuse` was 20 or 21, which is referred to as `landuse2021` wherever necessary. As already implied by the figure 2, the counties Glasgow City, Renfrewshire and Inverclyde had high concentration values irrespective of the type of land.

An Ordinary Least Squares model was implemented to remove the linear trend by the recorded co-variates. A forward variable selection strategy was used to select the variables and build the OLS model. 5% significance level was chosen in order to select the significant variables. The summary of the final chosen OLS model is given in the table 4. Using this, potential predictive co-variates for the spatial model were chosen.

Table 3: Summary of the OLS model

	Estimate	P-value
Intercept	12.1637851	5.86×10^{-12}
Y	-0.0000081	1.97×10^{-4}
Elevation	-0.0043945	2×10^{-16}
MRRTF	0.0800106	7.02×10^{-9}
Population	0.0000901	2.87×10^{-15}
High concentration county	0.1728052	1.22×10^{-5}
20 or 21 Landuse	0.0914153	1.46×10^{-3}
Adjusted R-squared	0.135	

Residual analysis, via diagnostic plots allowed to confirm that the model assumptions associated with non-spatial model are satisfied. The diagnostic plots are provided in the plot 5.

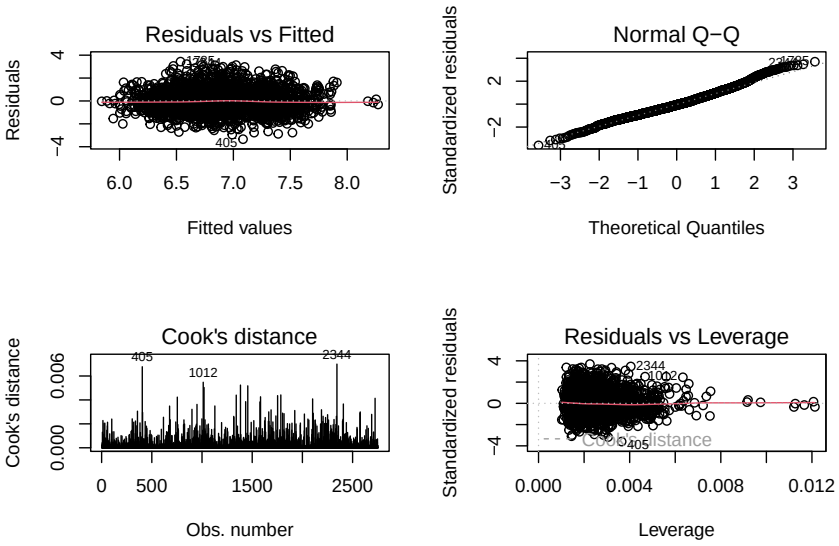


Figure 5: Diagnostic Plots for the fitted OLS Model.

From the plots, it is seen that the residuals are evenly spread out around the value zero, do not exhibit any pattern to indicate non-linearity or non-constant variance, and are roughly normally distributed as seen from the Q-Q plot. The cook's distance for all observations are also less than 1 indicating none of them are potential outliers. The independence of the errors will be assessed with the binned empirical semi-variogram.

The plot of residuals against co-ordinates in the figure 6 suggests no apparent spatial trend in the East-West or North-South direction. The first of the four plots uses quantiles to color-code the residuals, using which one could see some clusters in data region. The density plot reveals the normality of residuals which was shown in figure 5 Q-Q plot.

The short term spatial auto-correlation is assessed by the figure 7. The semi-variance at euclidean distance less than ~5000 lies outside the Monte-Carlo envelope suggesting very short term correlation, which needed

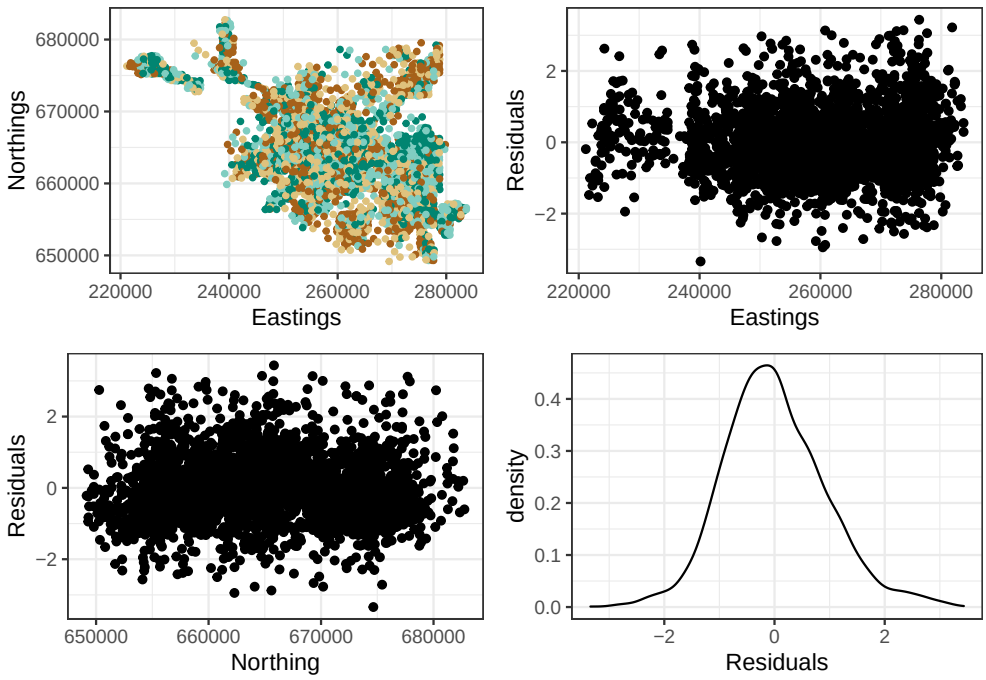


Figure 6: Residual Analysis for Clusters and Trends.

to be addressed. Thus the OLS model is adjusted accordingly to incorporate that information. Again, for short distances, the directional variogram for each direction seems quite similar, thus validating the isotropy assumption.

3.3 Effects of Co-variates on Spatial Pattern

Following the evidence of short-term spatial correlation among the values of concentration, a spatial model with exponential correlation structure was fitted at first. The choice of co-variates to be used was decided from the significant variables in the OLS model and the choice of correlation structure was arbitrary, but is almost always the default choice in spatial models.

Table 4: Spatial model summary with variables from OLS model.(* denotes significant variables)

	Estimate	Coefficient Interpretation
Intercept	11.5711460	
Y	-0.0000071	1000m increase implies 0.0071 decrease in log2 Pb.
*Elevation	-0.0049330	100m increase implies 0.005 decrease in
*MRRTF	0.0708499	1 unit increase implies 0.071 increase in
*Population	0.0000677	1000/sq.km. increase implies 0.068 increase in log2 Pb
*Landuse category 20 or 21	0.0854012	0.085 increase in log2 Pb if landuse 20 or 21
High Concentration County	0.0492582	0.005 increase in log2 Pb if high concentration county.
AIC Spatial Model	7195.815	
AIC OLS Model	7424.945	

A range of models with different correlation structures were used to choose the adequate model. After accounting for the spatial auto-correlation, the variables `Northing` and `High Concentration County` become insignificant at 5% significance level i.e, the effect of these variables dissipates after correlation is added to

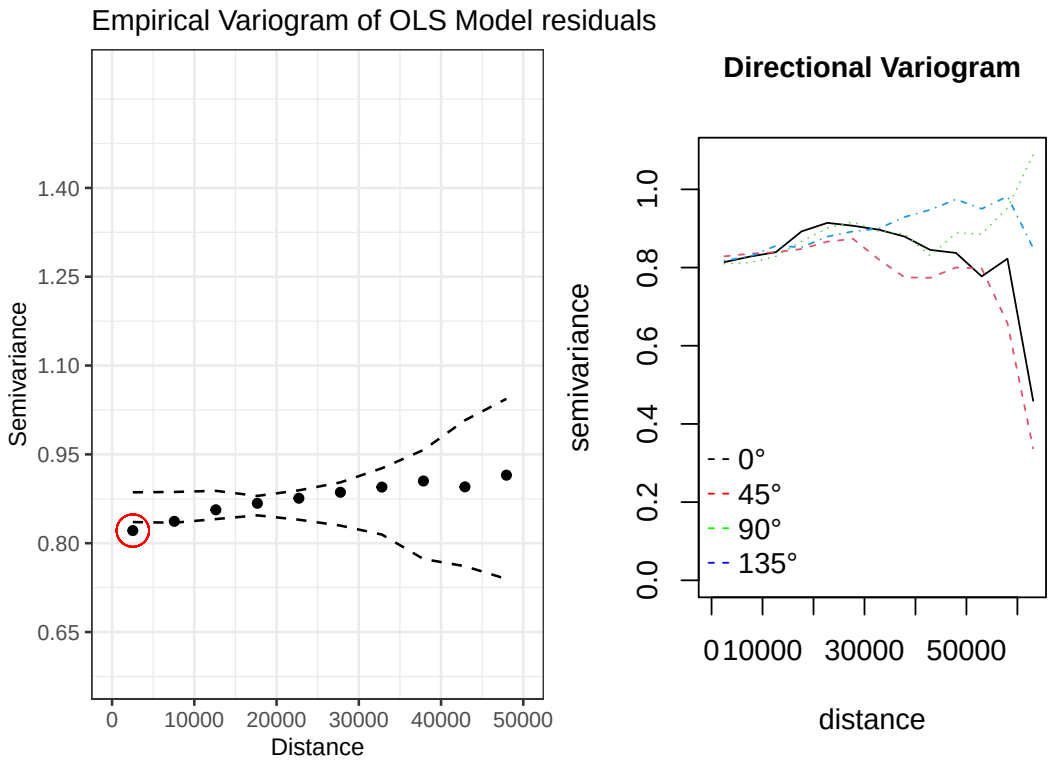


Figure 7: Empirical Variogram and Directional variogram to assess .

the model and were therefore removed from the spatial model. This effect was similar across the different models that were compared. The model adequacy criterion chosen was AIC, and as such, it was lowest for the model with exponential correlation structure, with a value of 7195.815. Comparing that with the AIC value of the OLS model given by 7424.945, it was seen that the spatial model is indeed a better fit to the data. The smooth variation (given by partial sill σ^2), random variation or noise (given by τ^2), and the correlation distance (given by ϕ), as estimated by the model, were 0.297, 0.708, and 5999.997 respectively. Nugget effect is higher than partial sill, which when coupled with the fact that the range parameter is low compared to the full range of distances (as shown in the variogram figure 7), suggested the presence of weak spatial auto-correlation, which nevertheless needed to be accounted for.

3.4 Outlier Analysis

As already stated, there were a total of 30 outliers which were labelled by the 3-standard deviation rule. There were no unusual values for co-variates that differed distinctively from the ones in the non-outlier case. The topographical indices were well within range and no interaction of co-variates explained these atypical values. However, the distribution of number of outliers across `Landuse` categories gives some direction, as provided in the plot 8. The importance of the categories' context is hence justified.

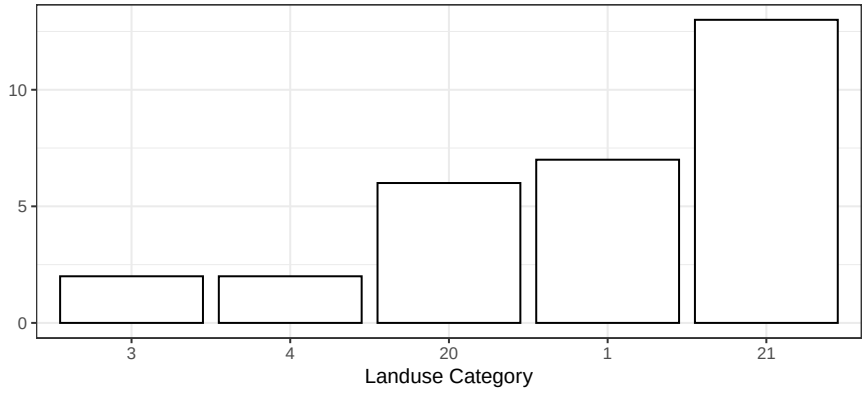


Figure 8: Number of Outliers across Landuse categories.

4 Research Conclusion and Discussions

The exploratory analysis along with the help of fitted models suggested that the Lead concentration had no discernible trends in North-South or East-West direction. Glasgow City council had the highest average level of Pb concentration, as was expected given the rich industrial heritage of the council. Landuse category 20 and 21 had high level of concentration irrespective of the councils. Some clusters of contamination were located in the maps. Glasgow City council had the most clusters which were easily visible in the concentration map. Most of the outliers were again from the Landuse category 20 and 21.

Evidences of correlation among the values of concentration were found, but the degree of correlation was deemed to be weak. There was a lot of unexplained variation in the data that was not covered by the models, possibly because of the low predictive ability of the sampled co-variates. Some of the major indicators of soil contamination such as pH value, loss of ignition etc were not available, the presence of which could have improved the explainability of the models. Nevertheless, there were certain variables found to be statistically significant (the most important of which is the indicator of `Landuse` category 20 or 21), but under the research context and their estimated effects, their practical significance is questionable.

This piece of independent research work could be extended in many segments to glean more information from the data. A map of prediction at un-sampled locations could be provided after improving the spatial model, possibly by including co-variates with better predictive power or by using a more advanced predictive algorithm altogether. A clustering algorithm that can include values of response along with easting and northing to identify clusters could be used to discern more subtle clusters.

5 References

Lanphear BP, Matte TD, Rogers J, Clickner RP, Dietz B, Bornschein RL, Succop P, Mahaffey KR, Dixon S, Galke W, Rabinowitz M, Farfel M, Rohde C, Schwartz J, Ashley P, Jacobs DE. The contribution of

lead-contaminated house dust and residential soil to children's blood lead levels. A pooled analysis of 12 epidemiologic studies. Environ Res. 1998 Oct;79(1):51-68. doi: 10.1006/enrs.1998.3859. PMID: 9756680.

Needleman H. L., Schell A., Bellinger D., Leviton A., Alred E. N. 1990 Jan. Long-term effects of exposure to low dose of Lead in childhood. The New England Journal of Medicine .

Fordyce, F.M.; Nice, S.E.; Lister, T.R.; O Dochartaigh, B.E.; Cooper, R.; Allen, M.; Ingham, M.; Gowing, C.; Vickers, B.P.; Scheib, A.. 2012 Urban soil geochemistry of Glasgow. Edinburgh, UK, British Geological Survey, 374pp. (OR/08/002) (Unpublished).

Urban Soil Geochemistry of Glasgow - Main report: Land Use Planning and Development Programme Open Report OR/08/002. https://nora.nerc.ac.uk/id/eprint/18009/

National Land Use Database: Land Use and Land Cover Classification version 4.4 2006 Feb. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/11493/144275.pdf

Ferraciolli M. A., Bocca F. F., Rodrigues L. H. A. 2019 Jun., Neglecting spatial autocorrelation causes underestimation of the error of sugarcane yield models .

Lee D. 2022, Stationarity and variograms, lecture notes block 2, Spatial Statistics 4H & 5M, University of Glasgow.

Lee D. 2022, Modelling Geostatistical Data, lecture notes block 3, Spatial statistics 4H & 5M, University of Glasgow.