The School of Mathematics

THE UNIVERSITY
of EDINBURGH

A Spatial Modeling Approach to Estimating Lead Contaminated Houses in Scotland

by

Yinying Xu, s1928699

August 2020

Supervised by
Prof. Finn Lindgren
Dr. Yiannis Papastathopoulos

Contents

Own Work Declaration

I confirm that the work of this MSc dissertation project is my own expect where otherwise declared.

Word count

4870 words (only applies to the main body of the text)

Executive summary

As lead contamination in water have a threat on public health in our daily life and the existence of remaining lead pipes is the main resource of lead in drinking water (Scottish Water 2020*b*), we try to develop a method of prediction for the proportion of houses that have lead contamination in Scotland.

In this case, we treated the total number of measurements detected lead as a response variable in our binomial regression model with a logit link function. We also created a 0/1 indicator for the absence or presence of lead contamination in the postcode. That binary variable follows a Bernoulli distribution so we built the logistic regression model with a logit link function to estimate lead contamination. In order to consider the variability of spatial indicators, such as regions, geographical coordinates, and the Water Operational area (WOA) that each sample belongs, we introduced the generalize linear mixed model and the generalized addictive model. Both model are ideal to analysis non-normal response from the exponential family distributions. The only difference is that the former treated region and WOA variables as random effects and the latter treated spatial coordinates as a smoother. Note that p parameter in the binomial model would be the estimate of the proportion of contaminated house in the postcode and can be transferred into the predictive probability of at least one samples detected lead (q) in the binary model. Thus, when assessing and comparing the prediction for different models with different distribution, we need to do the conversion of p into q; while operating the prediction, we need to convert q into p as p is approximately the proportion of houses contaminated in a single location. In this report, We used the Brier score that is the squared error between the prediction probability and the actual 0/1 classifications of lead contamination, and applied 5-fold cross-validation. To achieve better prediction, we want to minimize the average Brier score. Compared to modeling the binomial variable, the logistic regression models consistently produced inferior results for prediction. And the best prediction model is the binomial regression model with random effect of the name of WOA.

1 Introduction

Lead in drinking water has detrimental effects on health, which may cause respiratory, neurologic, digestive, cardiovascular and urinary diseases (Boskabady et al. 2018). Therefore it is of major concern of public health. However, high levels of this kind of toxic metal do not appear naturally in the drinking water supply. The issue arises when drinking water comes into contact with materials that contain or are made of lead. It is well recognized that lead pipes for delivering drinking water could increase the risk of lead poisoning (Rabin 2008). The widespread use of lead pipes in plumbing systems can be traced back to ancient Roman (Wikipedia contributors 2020*b*). Although the use of lead in pipes was gradually phased out during the 1960's and modern service pipes made of blue plastic took over, there are still some lead pipes that have not been removed from the UK plumbing network and remain in use, which contaminates the water supply (Rocha & Trujillo 2019).

Scottish Water, the primary water company in Scotland, has done substantial research and put a lot of effort into replacing the main supply pipes which are responsible for transporting water up to the boundary of the house (Scottish Water 2020*b*). However, there still exists quite a large number of communication pipes and supply pipes that are made of lead. This report aims to develop a method applied to identify areas throughout Scotland that are more likely to have lead-contaminated water supply given that one can access complete data. In other words, the end goal is to estimate the proportion of contaminated houses or the total number of that in Scotland. In terms of the small dataset given, estimating the proportion is more feasible approach to think about. If the sample of houses are representative in Scotland, we can obtain the estimated proportion of lead contamination, even without the total number of houses. We hope that, on the one hand, an increasing knowledge of the levels and location of lead can help Scottish Water to replace those lead pipes from the water distribution network and ensure citizens can reach lead free drinking water. On the other hand, a better understanding of lead contamination could raise public awareness of lead in drinking water and reduce their risk of exposure.

We will analyze 308 observations collected randomly from tap water in Scotland between 2010 and 2018. Each observation corresponds to a street postcode, while additional spatial information is also available including geographical co-ordinates of those samples based on British National Grid, the Water Operational area that each sample belongs and the corresponding geographical region. Our goal is to estimate the proportion of lead contamination in a certain area in Scotland by considering the effect of relevant factors including water and Scottish house quality determinants, and then assess and compare those predictions by using prediction scores.

2 Common sources of lead

Before we begin our analysis, let us give a comprehensive view of causes of lead in Scottish water supplies. We treated water do not have a significant level of lead naturally in the environment (Scottish Water 2020*a*). It is important to note of what are the most common sources of lead into Scottish water. Basically, they can be categorized into three disparate ways that are all related to materials that contain lead. Firstly, it is in connection with pipes that are applied to delivery water for daily use and these can be further classified into supply and communication pipes. Figure 1 shows the pipes involves in the public water supply. To be more specific, water runs from the water main underground and is distributed through communication pipes before entering into supply pipes to serve each household. Another main source is in household plumbing for pipes connection. It used to be done through the materials that contained or were made of lead. Therefore, lead particles being able to pose significant health risks from those materials used in household cooper pipe work dissolved into drinking water. It is also relevant with water tanks on the top of the house, which perform as the central system to allocate water for each household, that might be made of lead. In most cases, Scottish Water is responsible for the replacement of lead pipes from the water main in the street and communication pipes

up to the boundary of the house, whereas property owners are responsible for the section of the service pipe connected with the communication pipe and distributing water into the property itself, known as the supply pipe (Scottish Water 2020*b*). Thus, reducing lead cannot only rely on Scottish Water itself but also on each household.

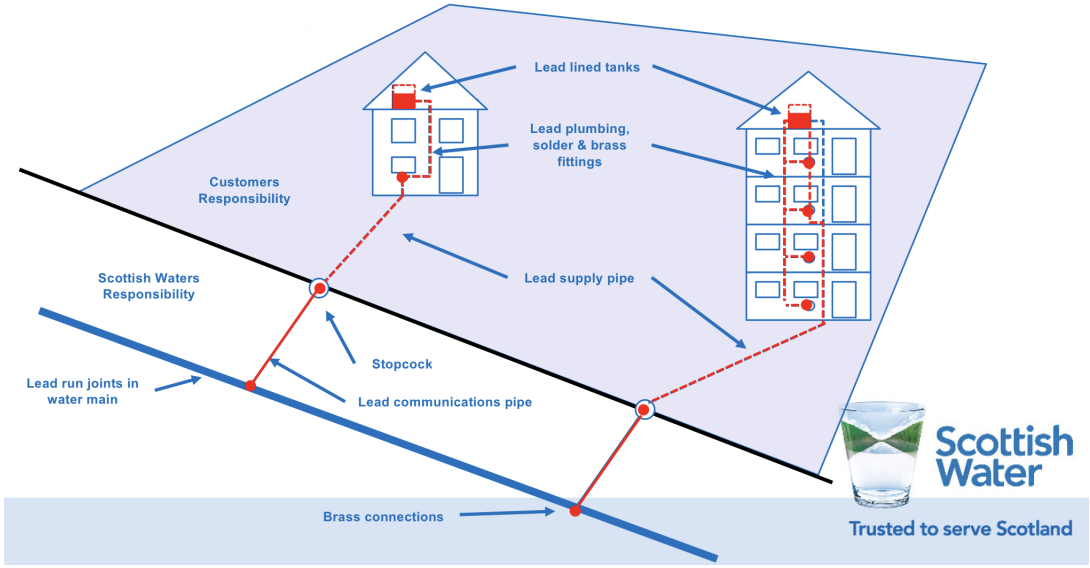

Figure 1: Scottish water supplies (image from Scottish Water)

3 Exploratory data analysis

The dataset used in this report contains information for 308 street postcodes where samples were taken from 2010 to 2018 in Scotland. This dataset provides water quality data collected from sampled households' taps and house quality data in that specific area.

3.1 Characteristics of water quality data

Figure 2 shows locations of sampled residential water quality tests and the level of lead contamination for each postcode in Scotland. Color responds to the average measured lead concentration of samples taken in the postcode (unite: μg per litre), while the size of circles represents the total number of samples that had over 1 μg of lead per litre tap water. We can see that raised lead measurements are highly geographically varied. It is worth noting that bringing the lead concentration levels down to zero is unachievable as many brass plumbing fittings used in copper pipework contain lead in low concentration level (Scottish Water 2020*b*). Therefore, throughout this report we emphasize that the street postcode with the average lead concentration over 1 μg per litre was contaminated by lead. Although the average value of lead measurement and the total number of samples detected lead over the threshold are two different ways to measure the level of lead contamination, this map also shows that those two variables are highly related to each other. We can observe that, if there were no measurements detecting lead more than 1 μg per litre, the average level of lead in water would be more likely to be lower than 1 μg per litre.

Except for the infrastructure work done to remove lead pipes that deliver drinking water, other control measures are taken to reduce lead in water, one of which being the orthophosphate dosing Scottish Water (2020*b*). Basically, phosphate is an anion derived from phosphorus and naturally can be found in water with a small amount. If lead pipes are treated with phosphate dosing which has the ability to effectively develop chemical layers of deposits around the pipes, it would be beneficial to prevent lead from dissolving into water and protect water supplies against contamination from this toxic heavy metal (Wikipedia contributors 2020*a*). Figure 3 shows most of points using a large amount of phosphate dosing are below the red dashed line

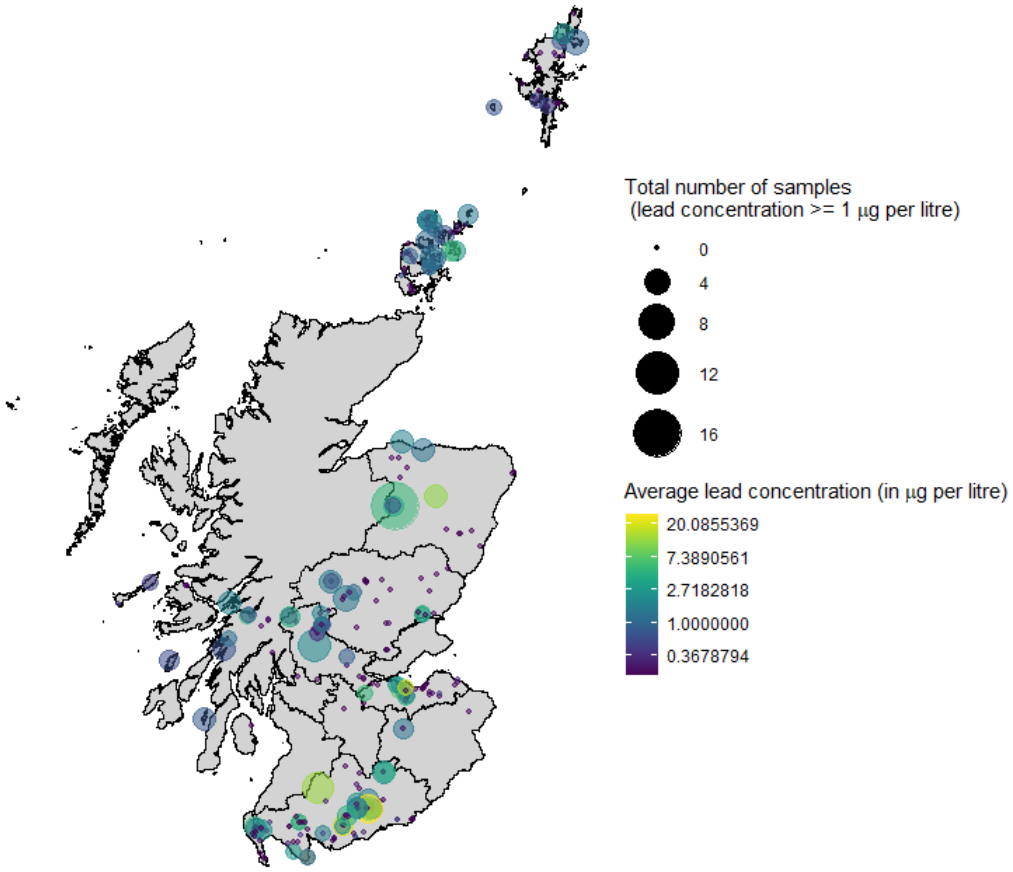

Figure 2: Locations of lead measurements taken in Scotland

at 1 μg per litre, while a few of them were still regarded as lead contamination. The right hand side of box plot also indicates that observations with at least one sample detected lead used the higher average amount of phosphate dosing. These implies that, in most cases, phosphate can act as a sufficient barrier to avoid lead leaking into the system. However, sometimes dozing put into the water treatment works was not so abundant that water could be sufficiently filtered and purified, which lead to failure to meet the water quality standard. Therefore, when the samples are taken, phosphate variable might mask lead pipes that are truly existing in that specific street postcode. Although it might be difficult for us to identify where lead contamination exactly is, it is still worth considering this predictor into our model to see whether that is significant or not and whether that improves predictions.

3.2 Characteristics of water quality data

Figure 4 shows the cases where old pipes have been replaced with new non-lead pipes, given a value of 1, are more likely to found more samples that have lead contamination than that where old pipes have not been replaced, given a value of 0. This is probably because the lead pipes were not necessarily replaced before the measurement was taken. That would cause difficulties for us to understand whether there is lead contamination now. To be more specific, this variable could be very useful if we want to predict other data points that were collected in the same way as that of the dataset we have. However, it does not mean that the model built for that can work as general prediction mode because the information about whether the lead pipes were replaced before or after the measurement was taken is missing. Even though having replaced pipes could act as a relatively strong indicator that lead would be found in the measurement in Figure 7, it might be because the lead was detected and then the pipes were replaced. As a

result of that, we will not consider this variable into our model for prediction.

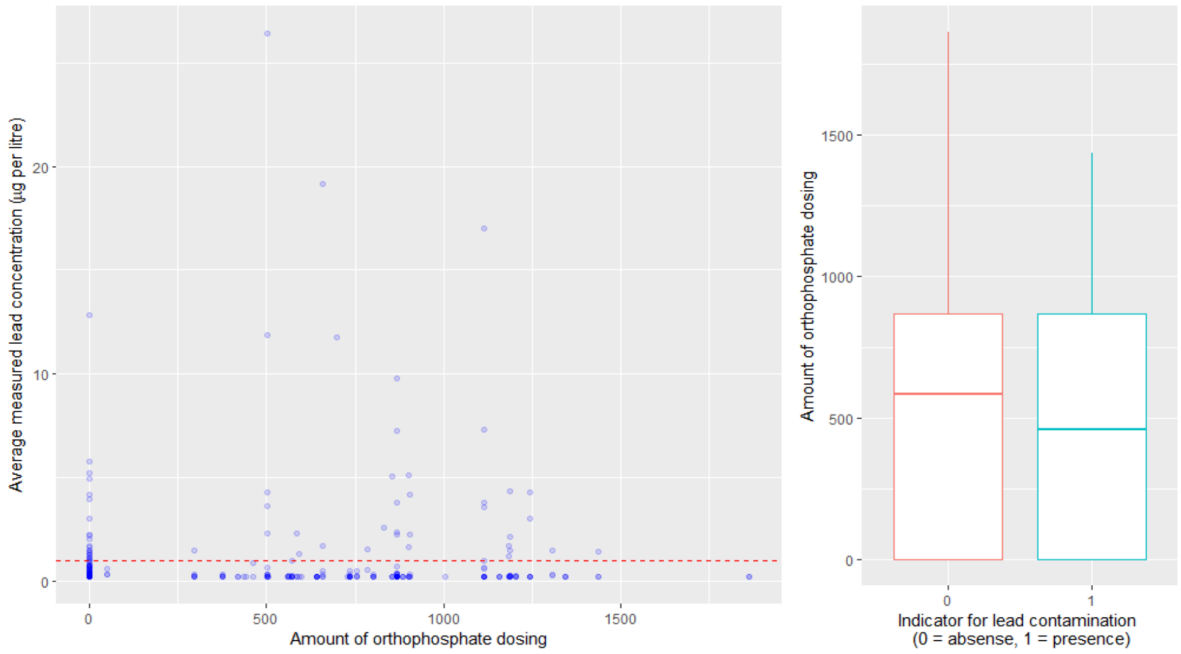

Figure 3: The relationship between the amount of orthophosphate dosing in drinking water at Water Treatment Zone and lead contamination

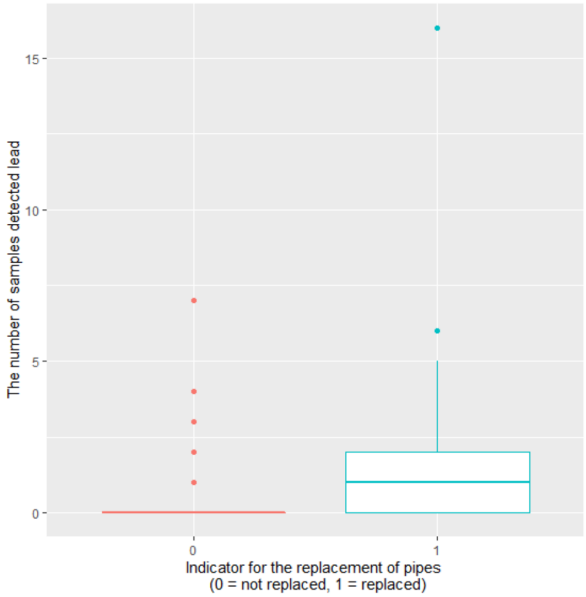

Figure 4: The boxplot of indication for the replacement of lead pipes against the number of samples detected lead

Significant concentrations of lead in the measurement is also relevant with the infrastructure built up to the 1960s when the use of lead pipes to deliver water in the UK plumbing network was prevalent. Those lead pipes were gradually got rid of during that decade and modern service pipes made of blue plastic took over. In 1969, lead was banned from use in plumbing, which was a further effort put into society to reduce lead in water (Scottish Water 2020*b*). In Figure 5, it seems that the larger proportion of houses in the postcode that were built after 1970, the lower average value of measured lead concentration. We can also see that the houses whose average building year was before 1969 in the postcode are more likely to have a higher level of lead on

average than that whose average building year after 1969. (See Figure 6) These imply that, if the house in the specific street postcode was built before lead became illegal, there is a chance that this area have lead pipework.

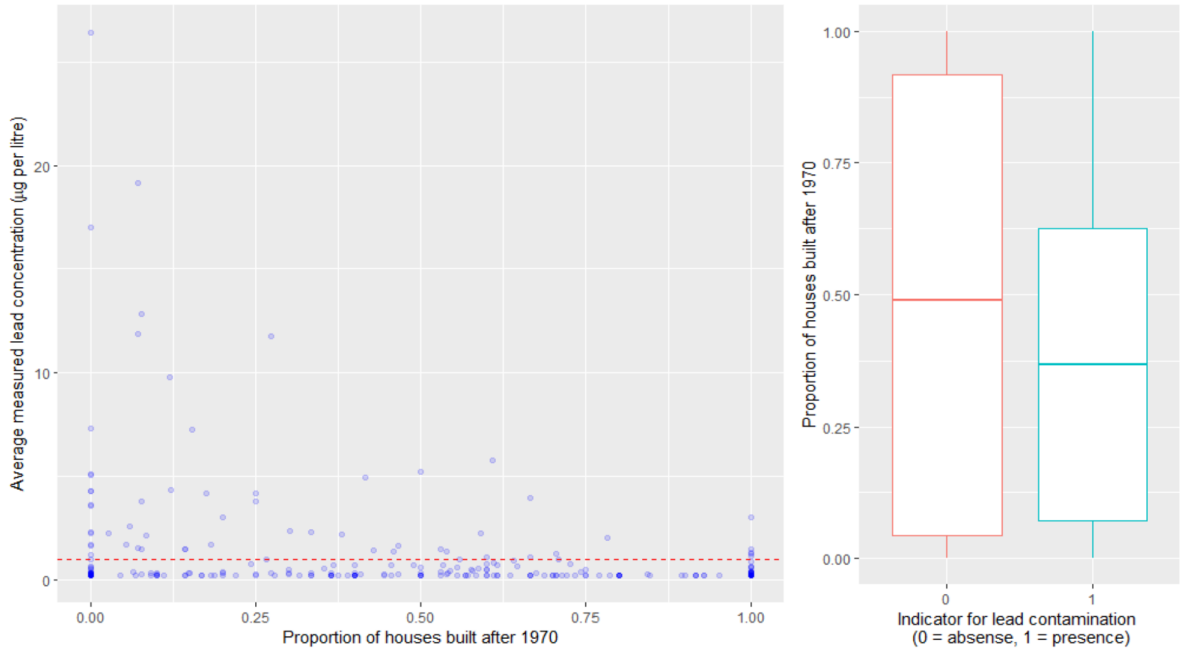

Figure 5: The relationship between the proportion of houses built after 1970 and lead contamination

There are more than one variable describing the location where samples were taken. For example, the geographical region that each postcode belongs to is given along with the indicator for how far the location of the postcode is from the city center or the remote rural. These are further classified into the corresponding Water Operational area (WOA). Moreover, there is information about Easting and Northing that are spatial coordinates of houses where the measurement was taken with respect to British National Grid, a specific coordinate reference system.

4 Method and model

4.1 Method

We applied the generalized linear model (GLM) first to estimate the parameters of house quality and water quality determinants in response to the dependent variable measuring the local lead contamination issue. Then we introduced the generalized linear mixed model (GLMM) and the generalized additive model (GAM) to take into account the variability of spatial locations. GLM is used to analyze the non-normal response from the exponential family distributions by applying link functions. To be more flexible, Gam can allow for smooth transformations of the response Faraway (2016). GLMM integrates the characteristics of GLM with linear mixed model including random effects (Bolker et al. 2009).

4.2 Selection of response variables and predictors

Before modeling, let us decide what our response variable is. In the dataset, there are two variables measuring the level of lead contamination for each postcode. One is the measured lead concentration on average, the other is the count of samples detected lead. Note that some of measurements were taken on the same household. This means that the count of samples detected lead is not exactly the number of lead-contaminated houses in the postcode, but we

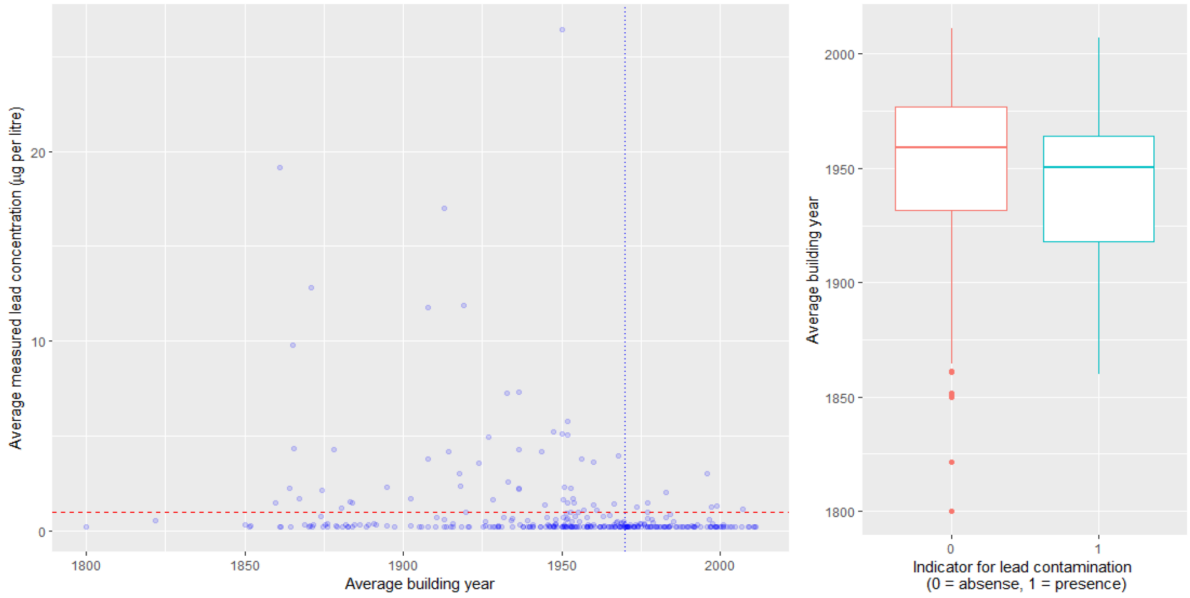

Figure 6: The relationship between the average building year for the houses in the street postcode and lead contamination

can assume that it is representative of the proportion of local lead contamination. As we aim at predicting the proportion of houses that have lead contamination, modeling the average of all measurements taken in the postcode directly cannot answer the real question. Even if we do the thresholding to get the binary version of the measured lead concentration on average first and then model that binary variable, the predicted probability that the lead concentration on average is greater than 1 μg per litre is not the same thing as the proportion of houses being lead contaminated. Finally, we tried to use the number of samples detected lead for each observation. We also created another variable by turning it into a binary variable to see whether the postcode was lead contaminated or not.

Based on the background of the issue in Sottish water supplies and initial data analysis, we will consider the following covariates in our full model for prediction:

- Phosphorus: amount of orthophosphate dosing used in the drinking water at Water Treatment Zone;
- BuiltAfter1970: Proportion of houses built after 1970 for each postcode;
- BuildYear: Average construction year for the houses in the postcode;
- MeanHouseAge: Average time, in years between the houses were built, and the time of the measurement;
- UrbanRural: Numerical indicator (between 1 and 8) for the distance from the city center.

However, Figure 7 indicates that all predictor variables and both response variables have a weak relationship, and a correlation of –1 between MeanHouseAge and BuildYear variables indicates a perfect negative correlation. Due to the goal that our model is built for prediction, we can include more covariates than otherwise we do, which might get better prediction score.

4.3 Binomial regression model

Let LeadPresence denote the number of samples detected lead over 1 μg per litre. Suppose the response variable $LeadPresence_i$ for i = 1,. . . ,n follows the binomial distribution with parameters $n \in N$ and $p \in [0, 1]$, which can be defined as

$$LeadPresence_i \sim Bin(n_i, p_i)\,,$$

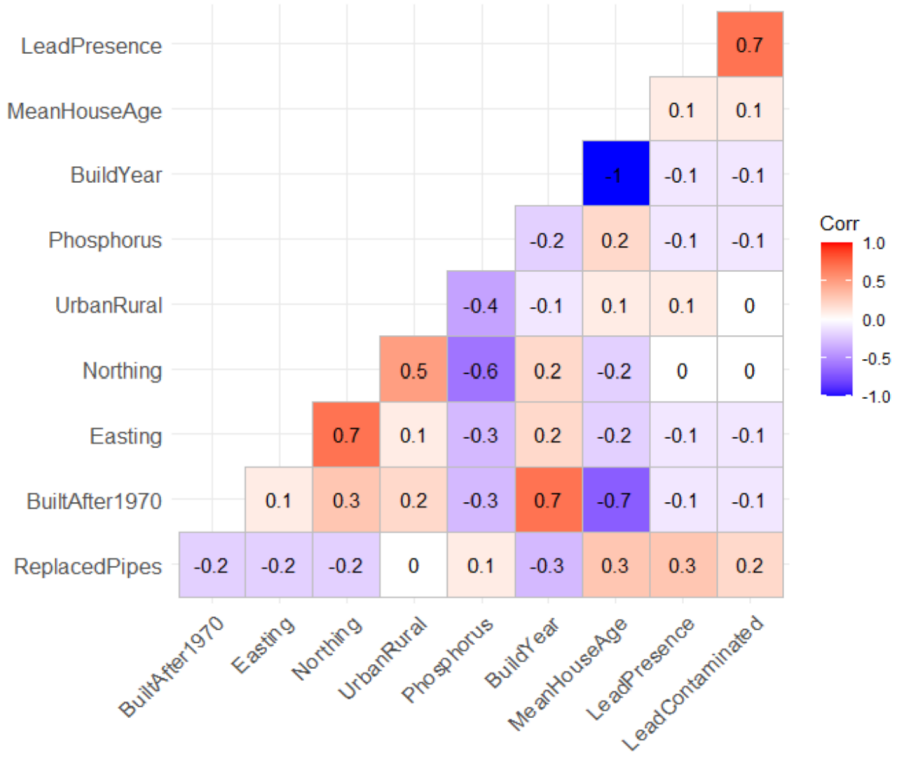

Figure 7: The correlation matrix

where n_i is the total number of measurements taken in the postcode i, p_i is the probability that any individual measurement shows lead contamination in the postcode i. We further assume that $LeadPresence_i$ are independent and households in the postcode were randomly tested for each measurement. For binomial response data, we can express it as a two-column matrix with the first column representing the number of successes $LeadPresence_i$ and the second column the number of failures $n_i - LeadPresence_i$ (Faraway 2016). In this case, a detected lead contamination incident is a "success" so that

$$P(LeadPresence_i = y_i) = \binom{n_i}{y_i} \cdot p_i^{y_i}(1-p_i)^{n_i - y_i}.$$

Then we construct a linear predictor:

$$\eta_i = X\beta,$$

where $\eta_i = log(p_i/(1-p_i))$ is a logistic link function and β's are regression coefficients. The expected number of samples detected lead would be

$$E(LeadPresence_i) = n_i p_i,$$

which can be rearrange as

$$p_i = E(LeadPresence_i)/n_i.$$

Therefore, if the samples are hopefully representative, the predicted probability p_i can be approximately treated as the proportion of houses that have lead contamination in the postcode i.

When fitting the full binomial model, we found that MeanHouseAge and BuildYear variables have the same values of the coefficients. That seems reasonable because MeanHouseAge variable can be recognized as the difference between BuildYear variable and the average time of water sampling in years, which means that MeanHouseAge variable depends on BuildYear variable but the former contains more information than the latter. Therefore, we decided to drop BuildYear

variable and remain MeanHouseAge variable for the following models. Thus, let the final GLM model with a binomial distribution denote Model 1, which can be expressed as

$$log(\frac{p_i}{1-p_i}) = \beta_o + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i.$$

4.4 Logistic regression model

Let LeadContaminated denote the binary version of LeadPresence, which can be written as

$$LeadContaminated = \begin{cases} 1, & \text{if } LeadPresence_i > 0 \\ 0, & \text{otherwise} \end{cases}$$

Suppose the response variable $LeadContaminated_i$ for i = 1,. . . ,n follows a Bernoulli distribution with parameters $q \in [0, 1]$, which can be defined as

$$LeadPresence_i > 0 \sim Bin(1, q_i)\,,$$

where q_i is the probability that at least one sample have lead contamination in the postcode i. Then we can write

$$P(LeadContaminated_i = 1) = q_i.$$

As the binomial response data, we can also construct a linear predictor:

$$log(\frac{q_i}{1-q_i}) = \beta_o + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i.$$

where β's are regression coefficients. We denote the binary model to be Model 2. The difference between using the binomial response matrix and the binary variable is that the former is modeling whether a given house in the postcode has lead contamination, and the latter indicates whether the postcode has any lead contamination.

It is important to note that, if we are modeling a binary variable, we should take into consideration how many possible values of LeadPresence variable and the total number of samples taken in the postcode they could be. This is because it is not the same number of measurements for each postcode. To be more specific, if there is the case where the only one measurement was taken in the postcode, then we are modeling the probability that the given measurement shows lead contamination. However, if the total number of measurements was greater than one and over one samples detected lead in the postcode, it means that at least one of all measurement was contaminated so the predicted probability q_i is not the same probability as an individual one being contaminated. Under assumptions that each sample was independent and was taken randomly in the postcode, we can convert that probability q_i into p_i. However, the issue with turning LeadPresence variable into a binary variable and the need to do that conversion of the predicted probability depend on a certain occasion. (We will talk it later)

As mentioned above, p_i is the probability of any individual measurement being contaminated in postcode i so $1 - p_i$ is the probability that a single sample is not contaminated. Then the probability that all of the measurements for that observation did not show lead contamination can be written as

$$P(all\ measurements = 0) = (1 - p_i)^{n_i},$$

where n_i is the total number of measurement taken in the postcode i and 0 means there is no presence of lead contamination. Thus the probability that at least one sample we get is positive can be expressed as

$$q_i = P(at\ least\ one\ measurement = 0 \mid postcode_i) = 1 - (1 - p_i)^{n_i}.$$

Therefore, when predicting the proportion of houses being contaminated in a new location k with m measurements, the binary model does the prediction of q_k, rather than p_k, and then we

need to do the conversion to solve for p_k that is

$$p_k = 1 - (1 - q_k)^{1/m_k}.$$

in order to predict LeadPresence variable that follows a binomial distribution, $Bin(m, p)$. However, it is arguable whether the q parameter can be predicted by p, since q depends on n in the binary model, whereas p and n are set independently in the full Binomial model.

4.5 Spatial model

We assume that the local lead contamination issue, and the responses to different house conditions and water quality data could vary randomly across different sampled locations. Thus we take the geographic nature of data into consideration by treating spatial covariates (i.e. WOAName and Region variables) as random effects and applied GLMMs where the random effect parameter estimated is the standard deviation of those covariates. The method of Restricted maximum likelihood (REML), a variant that averages over some of the uncertainty in the fixed-effect parameters, was used to estimate the random-effect parameters, providing less biased result than maximum likelihood (Bolker et al. 2009). Alternatively, in the GAMs, we added a smooth function f of location variables, i.e. Easting and Northing, and the models were also fitted using REML smoothing parameter estimation. In this report, we used "gam" function from "mgcv" package in R to fit our spatial model as mgcv can fit not only GAMs but also simple GLMMs through a spline that is equivalent to a Gaussian independent and identically distributed (i.i.d.) random effect.

Finally, we built several spatial models in terms of presence-absence data and the binary variable respectively. For LeadPresence variables that follows a binomial distribution, i.e. $LeadPresence \sim Bin(n, p)$, we fitted three models

- Model 3:

$$log(\frac{p_i}{1-p_i}) = \beta_o + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i + u_{WOAName};$$

- Model 4:

$$log(\frac{p_i}{1-p_i}) = \beta_o + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i + u_{Region};$$

- Model 5:

$$log(\frac{p_i}{1-p_i}) = f(Easting, Northing) + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i,$$

where u's are Gaussian i.i.d. random effects i.e. $u \sim \mathcal{N}(0,\, \sigma_u^2)$, based on various spatial variables and f is a smooth function.

Similarly, for LeadContaminated variables that follows a Bernoulli distribution, Bin(1,q), we also fitted three model

- Model 6:

$$log(\frac{q_i}{1-q_i}) = \beta_o + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i + u_{WOAName};$$

- Model 7:

$$log(\frac{q_i}{1-q_i}) = \beta_o + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i + u_{Region};$$

- Model 8:

$$log(\frac{q_i}{1-q_i}) = f(Easting, Northing) + \beta_1 Phosphorus_i + \beta_2 BuiltAfter1970_i + \beta_3 MeanHouseAge_i + \beta_4 UrbanRural_i,$$

where u's are Gaussian i.i.d. random effects i.e. $u \sim \mathcal{N}(0,\, \sigma_u^2)$, based on various spatial variables and f is a smooth function.

5 Model selection and comparison

5.1 Choice of proper prediction score

In the context of the issue that we have lead contamination, either positive or negative, positive means that lead was detected whereas negative means the opposite. Clearly, the prediction of that is whether there was lead contamination or not and we can tabulate the results of those combination. Refer to the lecture note of prediction assessment methods, let z_i be the true 1/0 classifications of lead contamination, and let $\hat{z}_i$ hat be predictions. The contingency table created to summarize True/False Positive/Negative, known as confusion matrix, is

Table 1: Condusion Matrix

	Predicted Positive	Predicted Negative
Lead Positive	TP = $\sum_i z_i\hat{z}_i$	FN = $\sum_i z_i(1-\hat{z}_i)$
Lead Negative	FP = $\sum_i (1-z_i)\hat{z}_i$	TN = $\sum_i (1-z_i)(1-\hat{z}_i)$

Usually, we can construct this "confusion matrix" both for 1/0 predictions, which means that we convert the prediction $\hat{z}_i$ into 1 if the prediction probability $\hat{p}_i$ is greater than 0.5; Otherwise into 0. In order to assess the prediction, we can check FP and FN, and measure the proportions of those rate of the whole dataset. Ideally, the model could be better if we get smaller sum of FP and FN (called the Absolute Error score). In our cases, Let $\widehat{C}$ be the predicted number of samples being positive, and C be the actual number of that. Then the prediction probability of lead contamination $\hat{p}$ and the actual probability of that p would be

$$\hat{p} = \widehat{C}/n = \sum_i \hat{z}_i/n,$$
$$p = C/n = \sum_i z_i/n.$$

As the end goal is to estimate the proportion of houses that have lead contamination in the Scotland, we want to minimum the difference between $\hat{p}$ and p. But the issue arises when we set $\hat{z}_i$ equal to 1 if $\hat{p}_i > 0.5$. For example, if the prediction probabilities are optimal, $\hat{p}_i = p_i = P(z_i|data)$, and all $\hat{p}_i < 0.5$ when the positive contamination incident is rare, the estimated proportion would be 0 rather than $\sum_i p_i/n$, which would result in the biased estimate. Therefore, taking predicted probabilities and turning them into 0/1 variables for this case is not an appropriate way, which will underestimate the contamination proportion.

A better way to assess the prediction discussed in the lecture is to measure the expected squared error of the predicted number of lead contamination, known as the Brier score. This

is commonly used when the response variable follows a Bernoulli distribution, $Bin(1, p_F)$. It is important to note that, when assessing the prediction for the binomial model, we need to convert the predicted probability q into p and then compare their average Brier scores with the binary model's. Under the assumption that the samples are representative and independent, this score can be defined as

$$S_{Brier}(F, z) = (z - p_F)^2,$$

Where z = the true 0/1 classifications of lead contamination.

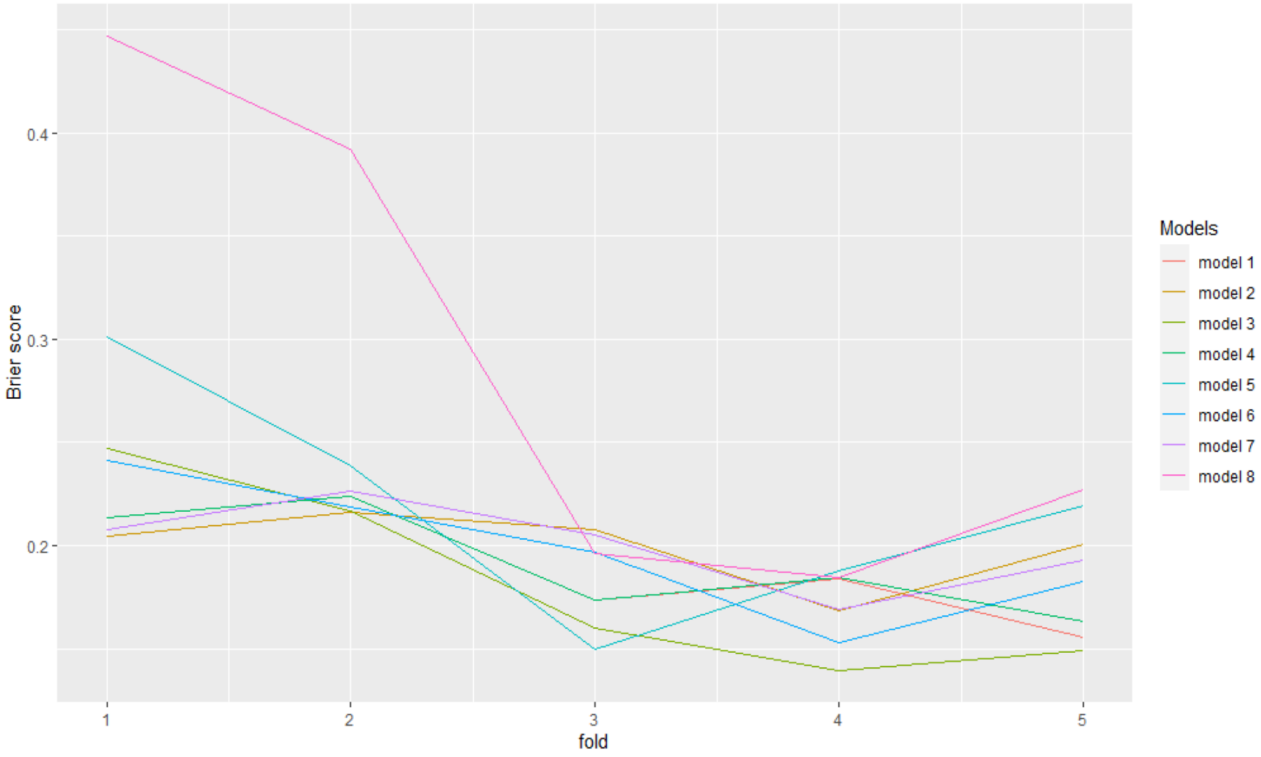

Figure 8: Plot of the average Brier score in each fold for each model

Table 2: The mean and variance of Brier score

	Score Mean	Score Variance
Model 1	0.1903	0.000794
Model 2	0.1996	0.000329
Model 3	0.1825	0.002196
Model 4	0.1918	0.000673
Model 5	0.2193	0.003232
Model 6	0.1985	0.001145
Model 7	0.2003	0.000442
Model 8	0.2894	0.014736

5.2 Cross validation

We applied cross-validation to assess how our predictive model built will generalize to an unknown dataset in real life and compare the performance between different models. In k-fold cross-validation, the N data points were randomly split into k subsets with the sample size of N/K. $K - 1$ subsets (known as the training dataset) are used to estimate the parameters of

the model, while the Kth subset (called the validation dataset or testing set) is used to assess the prediction. Then the cross-validation process is repeated K times so we can get the k Brier scores that are used to summarize the errors (Stone 1977). Due to the small dataset we have, we use the whole dataset to do 5-fold cross-validation and average the K results to produce an individual estimate of Brier scores for each model. The advantage of this method used to do model comparison is that we can use all data points for both training and validation, and each of the k subsets is used exactly once for validation. However, we would not assess the predictions on another dataset of unknown data, which might lead to the underestimation of the prediction error.

Figure 8 shows that Model 3 who have WOAName random effects consistently have the relative lower Brier score in most of folds, which means that model 3 is generally the best model for prediction in our report. The Table 2 also indicates that Model 3 has the minimum of the average Brier score at the point 0.1825 after the procedure of 5-fold cross-validation. In addition, compared to modeling the binary variable with different spatial variables, the binomial model can consistently produced the better prediction.

Table 3: Coefficients of Model 3

	Dependent variable:
	(LeadPresence, TotalPoints - LeadPresence)
BuiltAfter1970	-1.001^{**}
UrbanRural	-0.086
Phosphorus	-0.001
MeanHouseAge	0.007^{*}
Constant	-1.143
Observations	308
Adjusted R^2	0.368
Log Likelihood	-254.574
Note:	*p<0.1; **p<0.05; ***p<0.01

Table 4: Exponential transferred coefficients of Model 3

	Transferred Coefficients
BuiltAfter1970	0.3674
UrbanRural	0.9172
Phosphorus	0.9994
MeanHouseAge	1.0069
Constant	0.3187

5.3 The result of estimate and prediction

Here we give a brief description of the results for the best regression model, model . In Table 3, we found that both BuiltAfter1970 and MeanHouseAge variables are significant (at 5% and 10% level respectively), suggesting older houses are more likely to be contaminated. Also the variability of WOA location have a significant effect on the prediction proportion.

Before analyzing the effect of house and water quality determinants on the level of contamination, we used the exponential function to translate the regression coefficients as we applied logit transformation to the response. Table 4 shows the output of model with exponential coefficients. We can say that the odds of a single measurement that have lead contamination dramatically decrease by 63.26% with each additional proportion of houses that were built after 1970 in the postcode, but increase by 0.69% with each additional year in the average time between the average building year for the houses and the time of water sampling. For every unit increase in the location class of postcode i.e. as the location become closer to remote rural from city center, the odds of lead contamination, keeping BuiltAfter1970, Phosphorus and MeanHouseAge variables constant, reduce by a factor of 0.9172. In addition, Phosphorus variables have the lowest effect on the response. For each unit increases in the amount of orthophosphate dosing applied in drinking water, the odds of a single measurement that have lead contamination slightly decrease by 0.06%.

After doing the conversion of the predicted probability q into p, we assess the prediction on the whole dataset through the average Brier score at the point of 0.1240927 which is the Squared Error score for that predictive proportion of houses contaminated.

5.4 Model checking

In order to implement model checking for model, we looked at the diagnostic of deviance residuals against theoretical quantiles, it is acceptable that only a few points in the tail run out of the line of y = x. (QQ plot can be seen in github repository) This means that the data can be fitted well with binomial distribution.

6 Summary

We analyzed 308 observations collected randomly from the customer's tap water by Scottish Water from 2010 to 2018 with the goal of developing a method for the prediction proportion of houses that have lead contamination in Scotland. The dataset contains information about the house and water quality data for each observation.

We applied two main statistical methods: binomial regression; logistic regression. The former was directly modeling the number of measurements that was lead contaminated in the postcode that follow a binomial distribution, $Bin(n, p)$. The latter was modeling the binary variable, a 0/1 indicator for whether there was lead contamination, that follows a Bernoulli distribution, $Bin(1, q)$. Both of them used logit link functions and then we built spatial models through GLMMs and GAMs that take into account the variability of locations where samples were taken. Note that p would be the estimate of the proportion of contaminated house in the postcode and can be converted into the predictive probability of at least one samples detected lead (q). In order to assess and compare the prediction for different models, we introduced the Brier score that is the squared error between the prediction probability and the actual 0/1 classifications of lead contamination, and used 5-fold cross-validation. To achieve better prediction, we want to minimize the average Brier score. Compared to modeling the binomial variable, the logistic regression models consistently produced inferior results for prediction. Finally, we found that the binomial regression model with random effect of the name of WOA has the best performance for prediction.

A first limitation of the analysis is that the indicator for whether the older pipes were replaced might mask the true existence of lead pipes as some measurements might be taken before the

replacement of lead pipes, which is not ideal for prediction. However, this issue can be improved during the experimental design.

A second limitation is that whether the q can be fully transferred as p is questionable because q depends on n in the logistic regression model, whereas p and n are set independently in the Binomial regression model.

A third limitation is that, due to small dataset available, we use the full dataset to do cross-validation and cannot assess the prediction on the unknown data, which would result in the biased estimation for prediction.

References

Bolker, B. M., Brooks, M. E., Clark, C. J., Geange, S. W., Poulsen, J. R., Stevens, M. H. H. & White, J.-S. S. (2009), 'Generalized linear mixed models: a practical guide for ecology and evolution.', *Trends in Ecology & Evolution* **24**(3), 127–135.

Boskabady, M., Marefati, N., Farkhondeh, T., Shakeri, F., Farshbaf, A. & Boskabady, M. H. (2018), 'The effect of environmental lead exposure on human health and the contribution of inflammatory mechanisms, a review', *Environment International* **120**, 404 – 420.

Faraway, J. J. (2016), *Extending the Linear Model with R*, Chapman Hall/CRC Texts in Statistical Science, 2 edn, CRC Press.

Rabin, R. (2008), 'The lead industry and lead water pipes "a modest campaign"', *American Journal of Public Health* **98**(9), 1584–1592. PMID: 18633098.

Rocha, A. & Trujillo, K. A. (2019), 'Neurotoxicity of low-level lead exposure: History, mechanisms of action, and behavioral effects in humans and preclinical models', *NeuroToxicology* **73**, 58 – 80.

Scottish Water (2020*a*), 'Lead and your water'. `https://www.scottishwater.co.uk/your-home/your-water/lead-and-your-water`.

Scottish Water (2020*b*), 'What we are doing to reduce the amount of lead in your drinking water'. `https://www.scottishwater.co.uk/your-home/your-water/lead-and-your-water/what-we-are-doing-about-lead`.

Stone, M. (1977), 'An asymptotic equivalence of choice of model by cross-validation and akaike's criterion', *Journal of the Royal Statistical Society. Series B (Methodological)* **39**(1), 44–47.

Wikipedia contributors (2020*a*), 'Phosphate — Wikipedia, the free encyclopedia', `https://en.wikipedia.org/w/index.php?title=Phosphate&oldid=963891906`.

Wikipedia contributors (2020*b*), 'Plumbing — Wikipedia, the free encyclopedia', `https://en.wikipedia.org/w/index.php?title=Plumbing&oldid=963307560`.

Appendix

Detailed R code can be seen from the following link to a Github repository:`https://github.com/xu-echo/leadpipes2.git`