Estimating the number of homes in Scotland with internal lead piping

Gail Robertson, Amy Wilson, Yiannis Papastathopoulos, Margaret Graham, Lorna Eades, Efthalia Chatzisymeon, Chris Dent

1 Abstract

The University of Edinburgh undertook a project in partnership with the Drinking Water Quality Regulator (DWQR) to develop a statistical model to estimate the number of houses in Scotland which contain lead piping. The purpose of the statistical model was to provide enhanced quantification of the costs for replacing internal lead piping and thereby continue to ensure Scotland's high standard of drinking water supply. The project brought together expertise from the University's Schools of Mathematics, Chemistry, Geosciences and Engineering which worked in partnership with relevant Scottish Government departments.

2 Executive Summary

The aim of this project was to build a statistical model to provide an estimate of the number of houses in Scotland that are likely to contain lead piping or storage tanks and to identify postcodes which are likely to have increased numbers of houses with lead piping. This model would then be validated by carrying out two rounds of tap water sampling in postcodes throughout Scotland. To deliver this aim we first identified and collected appropriate datasets to be used in modelling, explored relationships between lead concentration and potential explanatory variables, made predictions of numbers of houses with lead piping within postcodes across Scotland using a best-fit model and validated the model using data collected in two rounds of tap water sampling. We sourced datasets containing lead concentration values from tap water samples taken from houses across Scotland as well as variables expected to explain lead concentration in tap water, such as house type and house age from Scottish Water, the Scottish Government's Scottish House Condition Survey (SHCS) and other sources. The scope of the project was limited to domestic properties connected to public water supply systems across Scotland. The specific objectives of the project were to:

- Obtain data from Scottish Water and the Scottish Government containing relevant information on internal lead piping, water quality, communication pipe replacement and house characteristics.
- Develop a statistical model using available data sources to determine which characteristics of individual houses are associated with internal piping/storage and from this determine which postcodes are likely to have more houses with internal lead piping.

- Examine model goodness-of-fit and use model estimates to develop tap water lead concentration sampling protocol for houses within 600 postcodes.
- Use tap water survey results to test accuracy of model predictions.
- Provide an estimate of number of houses with lead piping/storage tanks in Scotland and identify street postcodes with greater numbers of houses with lead piping.

Results: Our model showed that house age and whether or not lead communication pipes within a postcode had been replaced were important variables for predicting whether houses were likely to have internal lead piping. Applying our model to house age and communication pipe data throughout Scotland, we identified postcodes which are likely to have high numbers of houses with lead pipes. To validate our model we received tap water samples from a stratified sample of 326 postcodes and found that our model was able to predict the percentage of houses with lead pipes accurately for postcodes predicted to have large number of lead pipes, but provided less accurate estimates of percentages of houses with lead pipes in postcodes predicted to have low number of houses with lead pipes and postcodes in the Glasgow area.

Using our model to extrapolate the results across Scotland, our best estimate for the total number of houses in Scotland with lead piping was between 273,751 and 264,532 households (this varies depending on whether we include all postcodes with age data or only those in which >=10% houses have age data). We also estimate that the maximum number of households with lead piping (according to the upper level of a 95% prediction interval given by the model) lies between 1,397,912 and 1,232,554. Making predictions for all postcodes with explanatory variable data may lead to some inaccurate estimates if only a small proportion of houses in the postcode had age data. Only making predictions of number of houses with lead piping in postcodes in which more than 10% of houses had an age estimate leads to more accurate estimates of numbers of houses with lead pipes. Hence our estimate of 264,532 houses with lead pipes in Scotland is likely to be more accurate (with upper range of 1,232,554 according to 95% prediction intervals) (Figure 1).

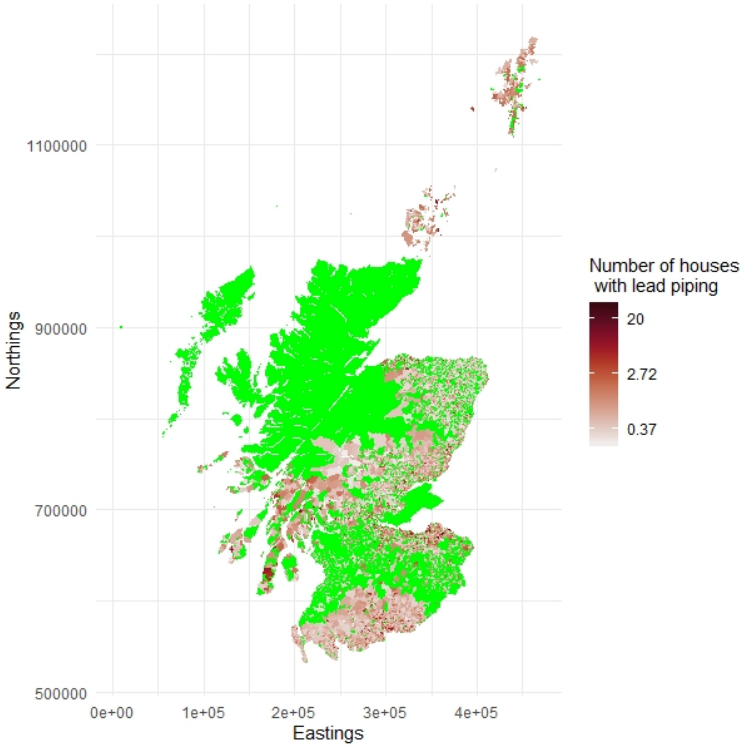

Figure 1: Map showing predicted numbers of houses with >=1 μg/l for 70,180 street postcodes including house age data. Green represents areas with no explanatory variable data

Key recommendations:

- **Data:** The model developed in this study was based on data that did not cover all postcodes in Scotland. In some postcodes the data available was limited and potentially biased (e.g. by the method of collection). There are other possible variables that may impact the probability that a household has lead pipes which could not be investigated due to lack of data. One key recommendation for improving the model and estimates described above is therefore to develop a comprehensive sampling scheme based on stratified sampling to ensure better coverage of samples across different postcodes and types of household.
- **Statistical modelling:** Improvements could be made to the model itself that may increase its accuracy. One such improvement is to account for spatial autocorrelation in the data. Another is to better capture uncertainty arising from the limited data.
- **Validation:** Model validation should be carried out using further tap water samples taken from a larger number of postcodes across Scotland, particularly those predicted to have lower percentage of houses with lead pipes.

Files provided with report:

- **Original model predictions:** Predicted number of houses with internal lead piping (lead concentration in tap water >=1 μg/l) per street postcode as estimated from original model (n=308) for all street postcodes with explanatory variable data ('Original model predictions and 95th quantiles all postcodes.csv') and only street postcodes in which house age data are available for >=10 houses ('Original model predictions and 95th quantiles selected postcodes.csv').
- **New model predictions:** Predicted number of houses with internal lead piping (lead concentration in tap water >=1 μg/l) per street postcode as estimated from refitted model including new data from sampling (n=386) for all street postcodes with explanatory variable data ('New model predictions all postcodes.csv') and only street postcodes in which house age data are available for >=10 houses ('New model predictions selected postcodes.csv').
- **Lead concentration data from both sampling rounds:** Lead concentration (μg/l) for each address sampled per postcode for both rounds of sampling (n=326). Mean number of houses with lead piping as predicted by the original model (n=308) is also included ('Lead conc data from sampled postcodes both rounds.csv').

3 Data

Data were initially available from Scottish Water from 2010 to 2018 and the Scottish Government's Scottish House Condition Survey (SHCS) from 2012 to 2017 (the final year for which data were available). Data available from Scottish Water included:

- Water quality data (lead concentration of tap water at addresses surveyed by Scottish Water using various sampling methods) (from 2010 to 2018)
- Communication pipe data (communication pipes carry water from water mains to the property boundary and are the responsibility of Scottish Water; includes number of pipes which have been replaced per postcode; material of pipes; length of pipes)
- Orthophosphate dosing levels and flow rates at each water treatment plant in 2018/2019
- Shapefiles showing street postcode and District Meter Area locations

Water quality and communication pipe data were available at the street postcode level, while orthophosphate dosing data were only available for water treatment plants which serve a large number of different postcodes.

Data available from the SHCS include the following information for each individual house (n = 3000) per year (2012 to 2017):

- Whether or not a surveyor had reported the presence of internal lead piping based on examination of plumbing visible in the house
- House condition; house type (detached, semi-detached, terraced, flat); age of house (years)

Data on house location are available at the street postcode level.

Some Scottish Water datasets provided data at different spatial scales to the SHCS data (street postcode level vs local authority or district postcode level). To link data from Scottish Water and SHCS for inclusion in the same model, we merged datasets together. On occasions where datasets to be merged included data at different spatial scales (i.e. where one dataset included data at the street postcode level and another at the district postcode level), data at the larger spatial scale were repeated per row in datasets which included data at smaller spatial scales. In such cases we assume that variables with data at smaller spatial scales are homogeneous over larger regions. As both datasets were at street postcode level from 2012 to 2017 we took into account that fact that street postcodes change over time as old postcodes are discontinued and new postcodes are created. As our aim was to make predictions for postcodes as close to the present date as possible, we selected postcodes in both datasets that remained consistent from 2010 to 2018, excluding those which were discontinued during this time period and those which had been created in 2018. We fitted the model to street postcodes available in 2019.

Shapefiles displaying locations of postcode districts throughout Scotland are freely available online as are definitions of urban and rural areas in Scotland as defined by the Scottish Government on an eight point scale (Figure 2). Postcode data were available from the National Records of Scotland as well as shapefiles showing locations of postcodes throughout Scotland from 2010 to 2019 (https://www.nrscotland.gov.uk/statistics-and-data/geography/our-products/scottish-postcode-directory/archived-postcode-extract). Postcodes at the district level are displayed in Figure 3.

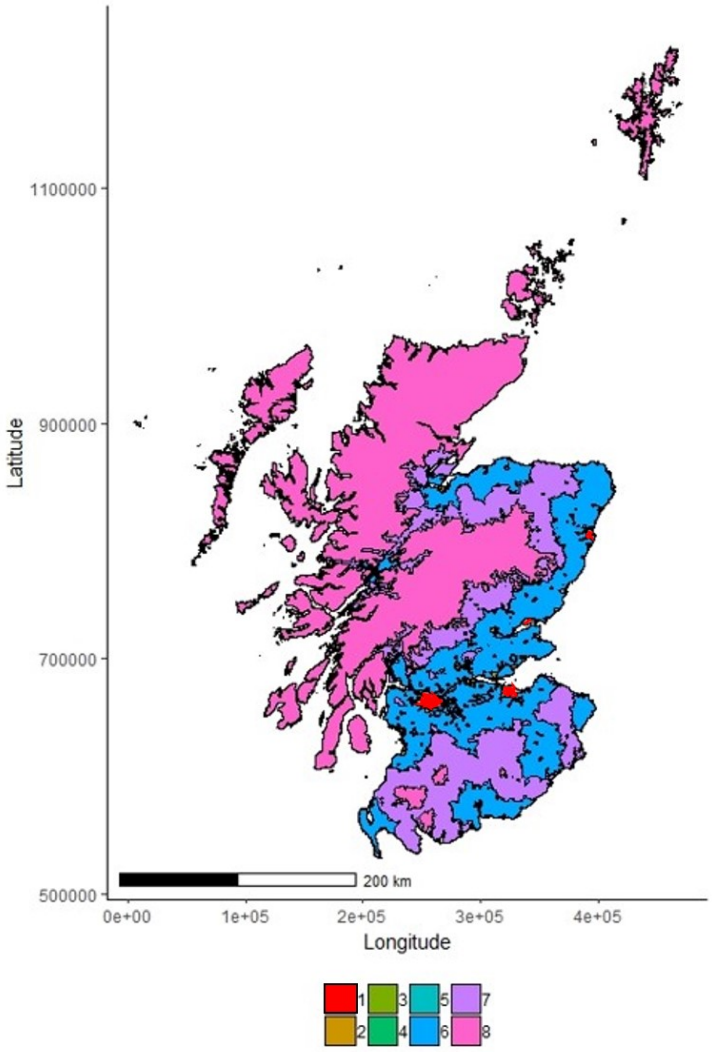

Figure 2: Locations of urban and rural areas in Scotland as defined by the Scottish Government Urban Rural Classification in 2016. The 8 fold classification definitions are as follows: 1 = Large Urban Areas (settlements of 125,000 or more people); 2 = Other Urban Areas (settlements of 10,000 to 124,999 people); 3 = Accessible Small Towns (settlements of 3000 and 9999 people and within 30 minutes drive of a settlement of 10,000 or more); 4 = Remote Small Towns (settlements of between 3000 and 9999 people and with a drive time of over 30 minutes to a settlement of 10,000 or more); 5 = Very Remote Small Towns (Settlements of 3000 and 9999 people and with a drive time of over 60 minutes to a settlement of 10,000 or more); 6 = Accessible Rural (Areas with a population of less than 3000 people, and within a 30 minute drive time of a settlement of 10,000 or more); 7 = Remote Rural (Areas with a population of less than 3000 people, and with a drive time of over 30 minutes but less than 60 minutes to a settlement of 10,000 or more); 8 = Very Remote Rural (Areas with a population of less than 3000 people, and with a drive time of over 60 minutes to a settlement of 10,000 or more)

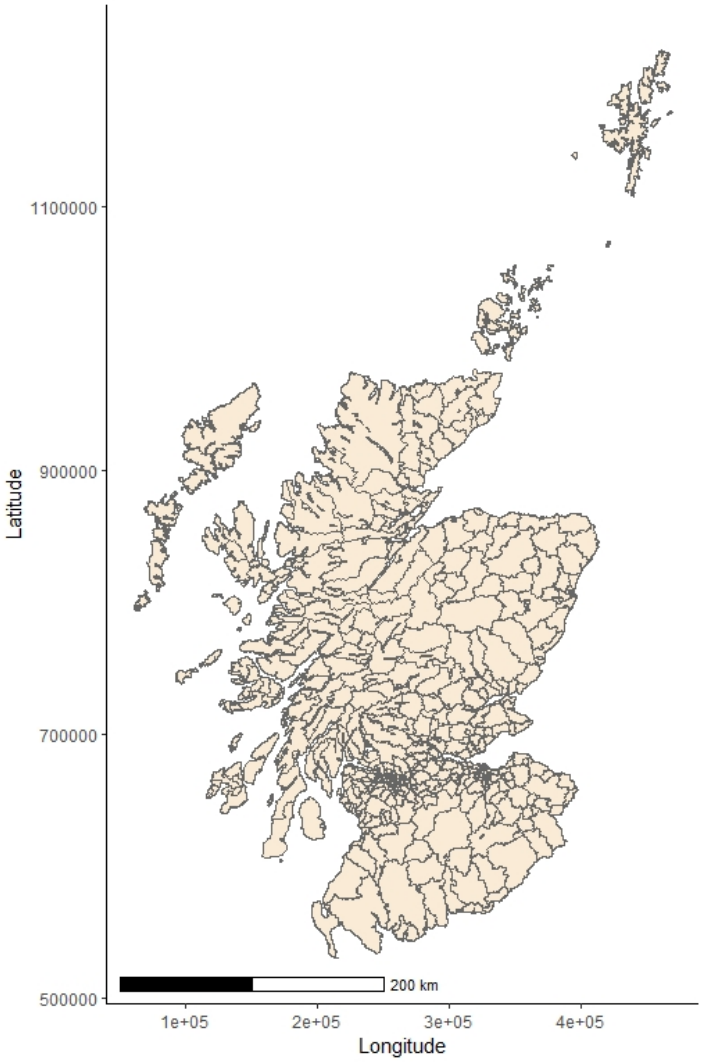

Figure 3: Locations of 444 postcode districts throughout Scotland

4 Bias in the data

There are various sources of bias in these datasets, most notably in variables representing whether or not a house had lead piping. The SHCS dataset records whether or not each respondents' house contained lead piping based on a survey of visible pipes carried out by a surveyor. Only 49 respondents recorded that their house had lead piping (out of a total of 16,800 respondents), possibly due to the fact that surveys were brief and only visible pipes were examined by the surveyor. Hence, responses given in the SHCS data may underestimate the number of houses with lead piping. Conversely, lead concentration of tap water recorded by Scottish Water during surveys may be overestimated due to some survey methods used to select houses, namely those in which customers contact Scottish Water to have their water tested. Customers who suspect their house may have lead piping are more likely to make such an enquiry leading to bias in the sampling

method and greater probability that lead concentration values will be higher from these houses. However, less than half of water quality sample data available from Scottish Water were collected via customer enquiry sampling (see Section 6.1) and our exploratory data analysis suggested that lead concentration levels in these houses were not significantly higher than those sampled using other methods.

5 Exploratory data analysis

5.1 Examine different survey methods used to collect lead concentration data

Lead concentration data was collected in various different ways depending on the methods used in each individual survey carried out by Scottish Water. In 'Scheduled' surveys tap water samples could be collected at any time of day, for 'Customer enquiries' water samples were only collected from taps in the first draw in the morning, whereas samples collected as part of 'Project' surveys could have been collected at any time of day, at first draw or after a defined flush period. The inherent differences in how samples were collected under each survey method may explain differences in lead concentration values among survey methods.
We intend to address the following questions regarding lead concentration values of water samples collected using different survey methods:

- How do lead concentration values vary depending on survey method used by Scottish Water?
- Which survey method is closest to a random survey? Should only randomly collected data be used in further analysis?
- Did houses sampled earlier in the day have higher lead concentration values? If so, we might expect this to be due to lead build up in pipes overnight and we may consider having to control for timing of sample in models.

Note that for data exploration we used as much of the available data as possible, in that we used Scottish Water water quality data from all available years (2010 to 2018) to explore as much variation in lead concentration as possible. When data were merged with SHCS data, we used Scottish Water water quality data from 2010 to 2018 to maximise the number of points available from both surveys. To do this we assume that survey results for lead concentration are comparable to SHCS data regardless of the year in which the survey was carried out. Table 1 shows the number of houses surveyed per year and summary statistics of lead concentration values per year. The figures in Table 1 show a decreasing trend in mean lead concentration values from 2010 to 2018. It is unclear what is causing this effect, whether it is due to the cumulative effect of orthophosphate dosing or whether over time more and more lead water distribution pipes have been removed from the network reducing lead concentration of tap water. Whatever the cause, it is important to note that lead concentration in sampled houses appears to change over time.

Two survey methods from Scottish Water water quality dataset 'Grid log' (n=40) and 'Resample' (n=3) were excluded from further exploratory analysis as houses surveyed via the 'Grid log' method did not have postcodes recorded at the street postcode level and resamples were duplicate samples taken from the same house. Figure 4 shows the distribution of lead concentration values collected using different types of tap

water surveys implemented by Scottish Water. The plot suggests that lead concentration of tap water is higher in houses surveyed during 'Scheduled' surveys and lower in 'Project' surveys. Table 2 and Figure 5 display distribution of lead concentration values collected via different survey methods.

Table 1: Number of houses surveyed and summary statistics for lead concentration (in μg/l) of tap water in houses surveyed each year by Scottish Water from 2010 to 2018 using all survey methods.

	2010	2011	2012	2013	2014	2015	2016	2017	2018
Number of houses	8285	8651	9568	8826	9272	4490	5273	6140	4984
Mean lead concentration	13.22	10.32	7.58	7.23	5.51	3.46	3.25	3.47	3.68
Standard deviation lead concentration	41.67	44.15	30.71	31.38	26.63	27.23	23.88	30.17	32.37
Median lead concentration	2.8	1.7	1.3	1.5	1.0	0.2	0.2	0.2	0.2

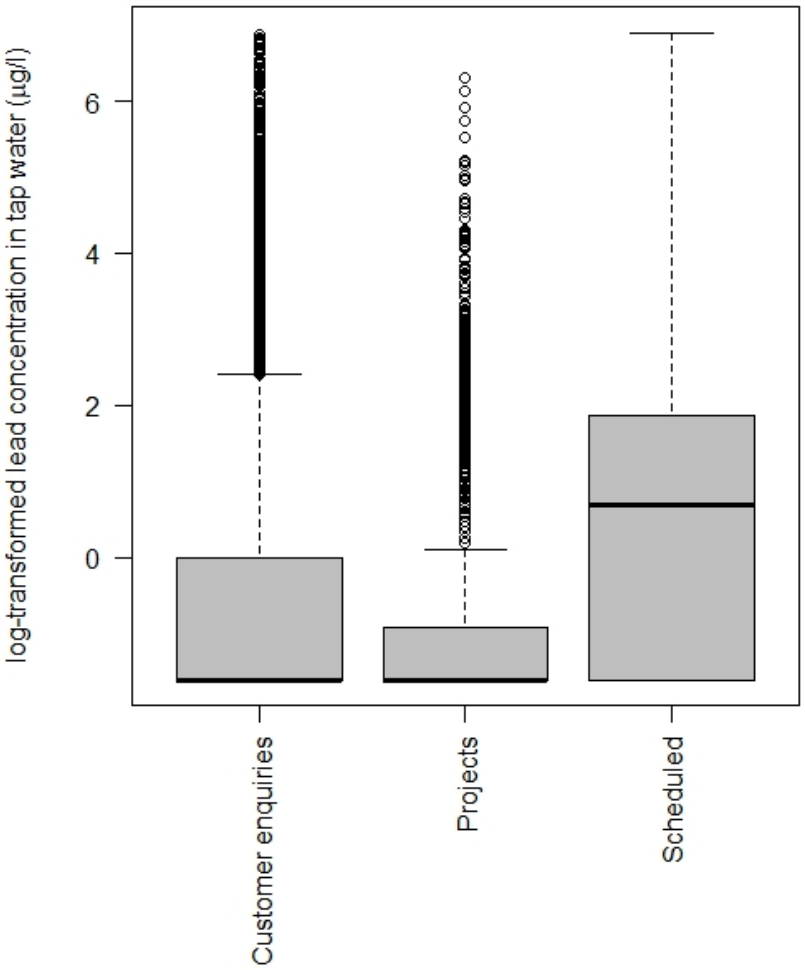

Figure 4: Boxplot showing log-transformed lead concentration of tap water in houses surveyed using different methods. The 'Customer enquiries' method refers to houses selected for sampling after customers contacted Scottish Water asking for samples to be taken; the 'Projects' method refers to houses included in a survey as part of a specific research project or investigation carried out by Scottish Water; the 'Scheduled' method involved Scottish Water surveying randomly selected houses every year as part of an annual survey process

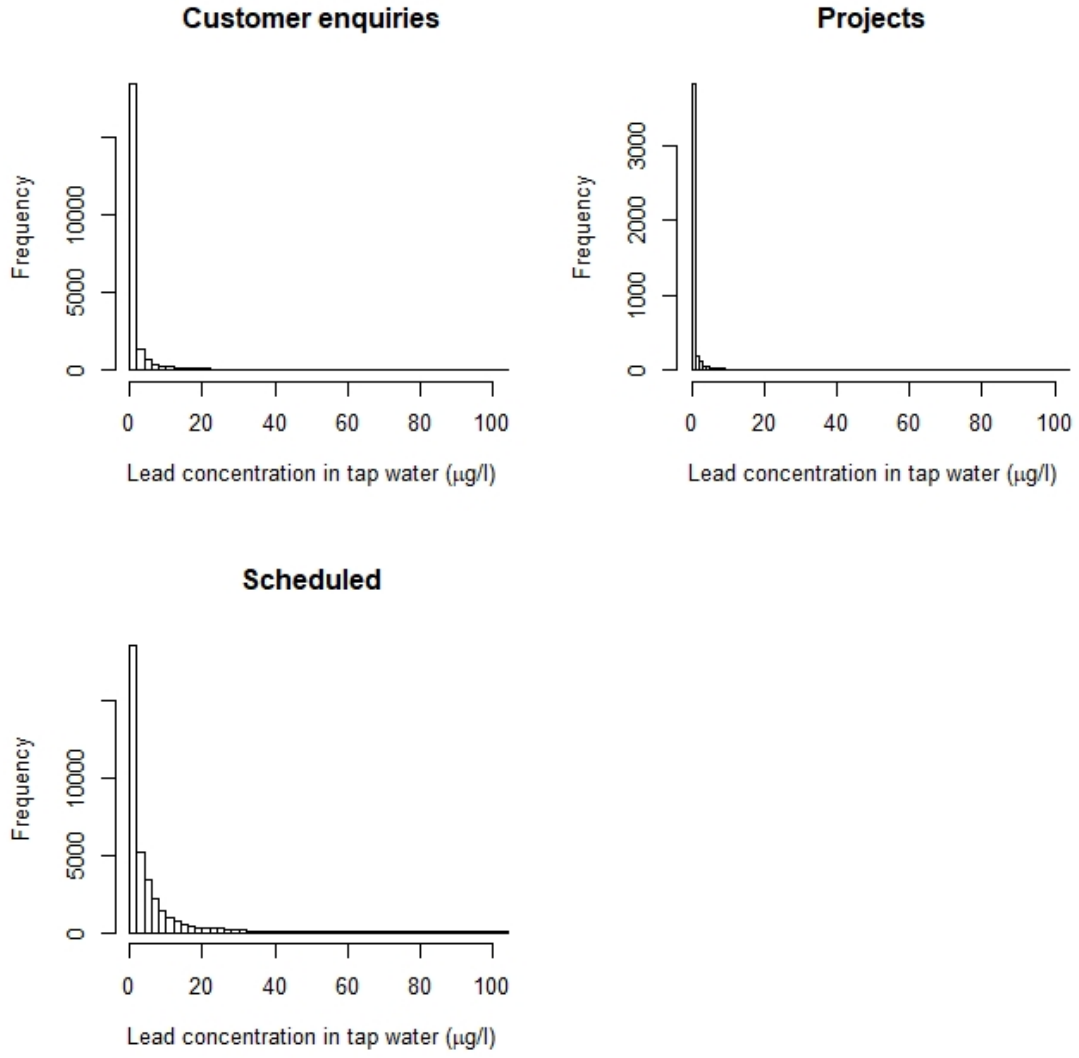

Figure 5: Histograms showing lead concentration of tap water in houses surveyed by Scottish Water using different methods. Note differences in values of y-axes

Table 2: Summary statistics of lead concentration values of tap water (in μg/l) in houses surveyed using different sampling methods.

	Survey method	n	Mean	Standard Deviation	Median	Min	Max
1	Customer enquiries	23387	5.04	36.11	0.20	0.20	962.00
2	Projects	4761	2.24	18.35	0.20	0.20	550.50
3	Scheduled	37341	8.79	33.10	2.00	0.20	994.80

We used two separate models to compare lead concentration values in houses surveyed using different survey methods implemented by Scottish Water. Due to the distribution of the data (see Figure 5) we used two models to represent different parts of the distribution: one in which the response variable (lead concentra-

tion) was separated into two groups (<1 or >=1 μg/l) and another modelling only lead concentration values over 1 μg/l. We used a threshold value of 1 μg/l beyond which a house is assumed to have internal lead piping and below which a house is assumed to be lead-free based on recommendations from Scottish Water and the Drinking Water Quality Regulator.

Model results suggest that houses without lead pipes (those with lead concentration values of <1 μg/l) were more represented in Project surveys than Customer enquiry surveys while Scheduled surveys had a greater representation of houses with lead pipes (see Table 3). For lead concentration values >=1 μg/l we found the same pattern (see Table 4). Scottish Water has stated that the Scheduled survey method randomly selected houses to be tested as part of an annual monitoring program, while use of Customer enquiries is likely to result in a biased lead concentration estimate as customers who suspect their house has lead piping are more likely to make an enquiry. However, our analysis suggests that houses sampled as part of Scheduled surveys had greater lead concentration values than those sampled via Customer Enquiries. Based on recommendations from Scottish Water about different sampling methods, further analysis was carried out using data collected as part of Scheduled surveys only.

Table 3: Results of a binomial generalised linear model comparing lead concentration values (where values are <1 (no lead) or >=1 (lead) μg/l) given by different sampling methods employed by Scottish Water. Projects and Scheduled values are given in comparison to data collected via Customer enquiries.

	Coefficient	St.error	z.value	p-value
Intercept	-1.06	0.02	-69.87	<0.001
Projects	-0.68	0.04	-15.27	<0.001
Scheduled	1.46	0.02	78.63	<0.001

Table 4: Results of a gamma-distributed generalised linear model comparing lead concentration values (where values are >=1 μg/l) given by different sampling methods employed by Scottish Water. Projects and Scheduled values are given in comparison to data collected via Customer enquiries.

	Coefficient	St.error	z.value	p-value
Intercept	1.58	0.02	103.11	<0.001
Projects	-0.20	0.05	-4.08	<0.001
Scheduled	0.23	0.02	13.11	<0.001

Figure 6 suggests that houses surveyed earlier in the day had higher lead concentrations than those surveyed later. We may need to consider this temporal pattern when identifying factors explaining differences in lead concentration between postcodes.

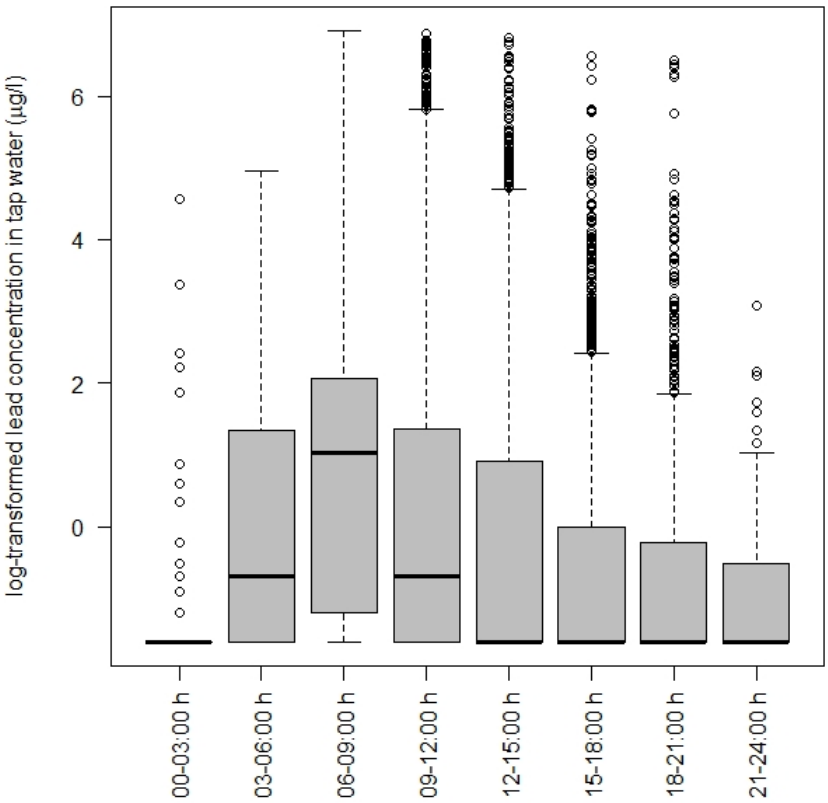

Figure 6: Boxplot showing log-transformed lead concentration of tap water in houses surveyed at different times of the day. Resampled houses (n=3) and Grid Log survey data (n = 40) are excluded

5.2 Examine the relationship between presence/absence of lead piping and lead concentration in tap water

Only 49 householders in total reported having lead piping according to the SHCS. To find out whether lead concentration data from Scottish Water surveys may be used to determine whether or not a house had lead piping we compared presence/absence of lead piping in houses in available street postcodes with lead concentration values (in μg/l) collected by Scottish Water.

Only 25 postcodes contained houses which reported presence of lead piping in the SHCS for which Scottish Water lead concentration data were also available.

We attempted to address the following questions:

- Is there a relationship between lead presence/absence and lead concentration?
- How strong is this relationship? Does it suggest that we can be justified in using lead concentration as a proxy for presence/absence of lead piping?

Table 5 shows numbers of unique responses given by respondents in the SHCS in regard to whether or not respondents have lead piping in their houses.

Table 5: Number of respondents in the Scottish House Condition Survey who reported having lead pipes in their house in the whole survey and those respondents in postcodes which also have lead concentration data

	Presence of lead pipes	No lead pipes
All survey data	49	16731
Survey data with water quality data	25	7942

Figure 7 shows that houses with internal lead piping were located in postcodes in which houses surveyed by Scottish Water also had higher lead concentration values of tap water. These include data for all street postcodes and no aggregation of lead concentration values per street postcode was carried out.

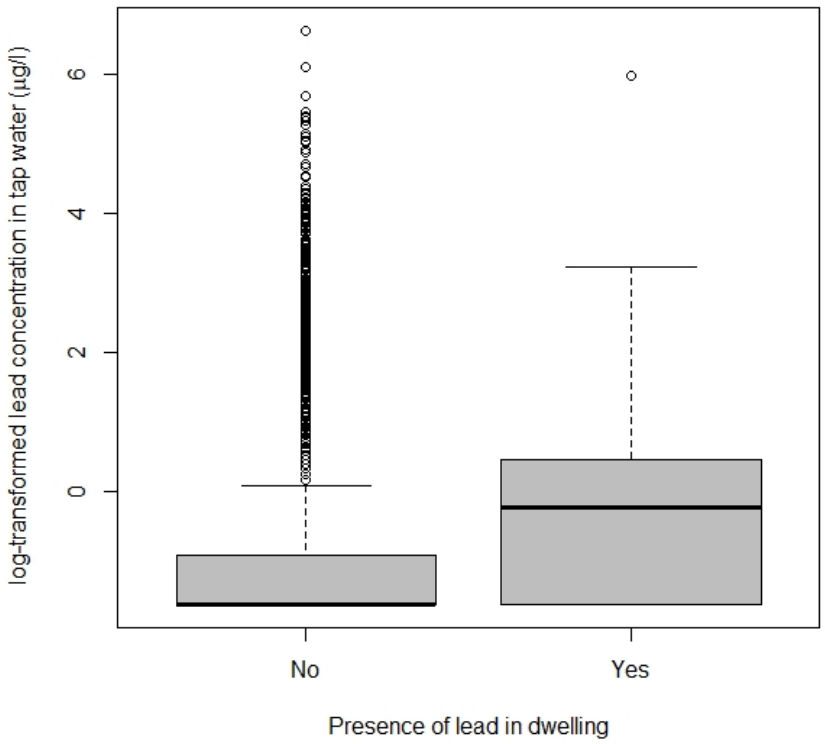

Figure 7: Boxplot showing log-transformed lead concentration of tap water in houses with different lead piping statuses (lead piping present n = 25, lead piping absent n = 7942). Resampled houses (n=3) and Grid Log survey data (n = 40) are excluded

We used a binomial generalised linear model to determine whether lead concentration of houses sampled in postcodes where houses reported having lead pipes was higher than in postcodes where houses did not report having lead pipes. Table 6 displays model results which suggest that although there was a significant difference in lead concentration values between postcodes with houses with and without lead pipes, the effective difference is very small. When this model was repeated using a dichotomous variable representing lead concentration (with threshold of 1 μg/l, where <1 and >=1 μg/l) were taken to represent absence and presence of lead piping respectively) we found similar results, but a larger effect size (i.e. the strength of the relationship between presence of lead piping and lead concentration threshold was greater than that describing the relationship between presence of lead piping and lead concentration) (Table 7). Note that these data include repeated values for street postcodes where one to several houses were surveyed for lead piping presence and lead concentration of tap water.

Table 6: Results of a binomial generalised linear model examining the relationship between presence/absence of lead pipes in houses within the same street postcodes and lead concentration values (in μg/l).

	Coefficient	Std error	z value	p-value
Intercept	-5.79	0.2	-28.57	<0.001
Lead concentration	0.005	0.002	2.72	0.01

Table 7: Results of a binomial generalised linear model examining the relationship between presence/absence of lead pipes in houses within the same street postcodes and lead concentration values (with threshold of 1 μg/l, where <1 and >=1 μg/l were taken to represent absence and presence of lead piping respectively).

	Coefficient	Std error	z value	p-value
Intercept	-6.16	0.27	-23.07	<0.001
Lead presence threshold	1.40	0.40	3.47	<0.001

5.3 Relationship between lead concentration of tap water and explanatory variables

Presuming that lead concentration data can be used as a proxy for presence/absence data we now explore relationships between lead concentration and house age (or year in which house was built); type of house; condition of house and year. In order to determine whether these relationships vary over space we created maps showing spatial variation in lead concentration around Scotland as well as the distribution of type of house and condition of house.

After merging the SHCS and water quality datasets we were left with 7977 rows of data (over 3453 unique postcodes) with both housing data and water quality data available. For some postcodes in this dataset only one or two houses were surveyed, whilst for other postcodes multiple houses were surveyed in same or different years. Figure 8 shows presence/absence of lead piping in surveyed houses vs year house was built and lead concentration of tap water in houses surveyed by Scottish Water in the same street postcode. Plots suggest that respondents that reported having lead piping in their house tended to live in older houses than those that did not report lead piping and lead concentration appeared to be higher in postcodes with older houses.

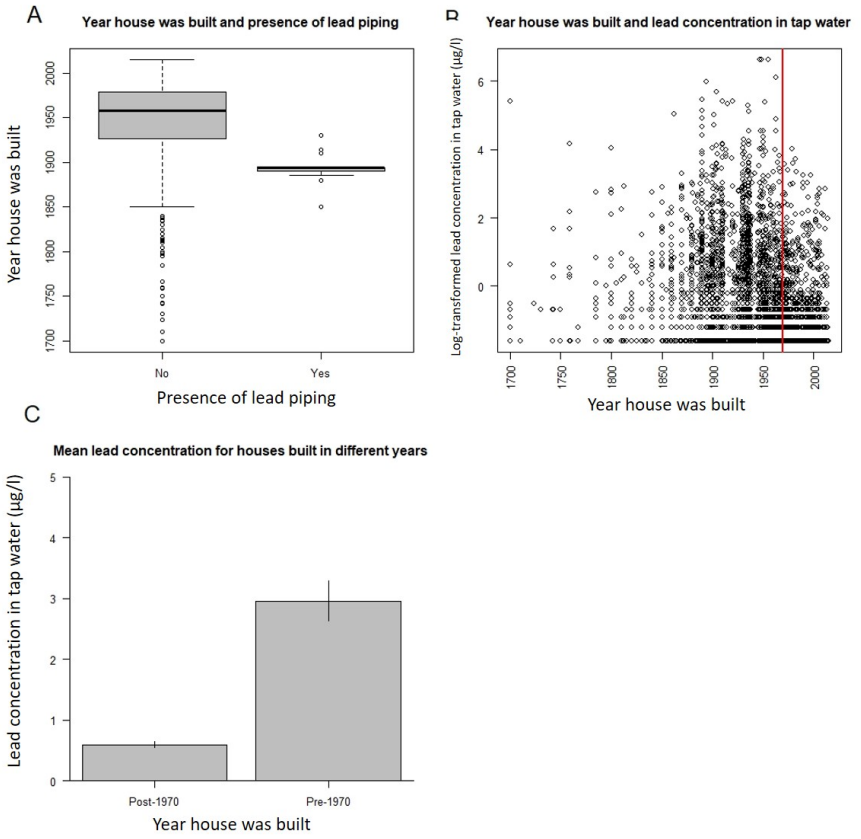

Figure 8: Plots exploring relationships between lead piping presence in houses/lead concentration of tap water and year house was built for A) House age and presence of lead piping (No n = 7942; Yes n =25); B) House age and log-transformed lead concentration in tap water; C) Mean and standard error lead concentration of houses built pre- and post-1970. Note that year house was built and lead concentration were not recorded for the same individual houses but for houses surveyed in the same postcode. Resampled houses (n=3) and Grid Log survey data (n = 40) are excluded

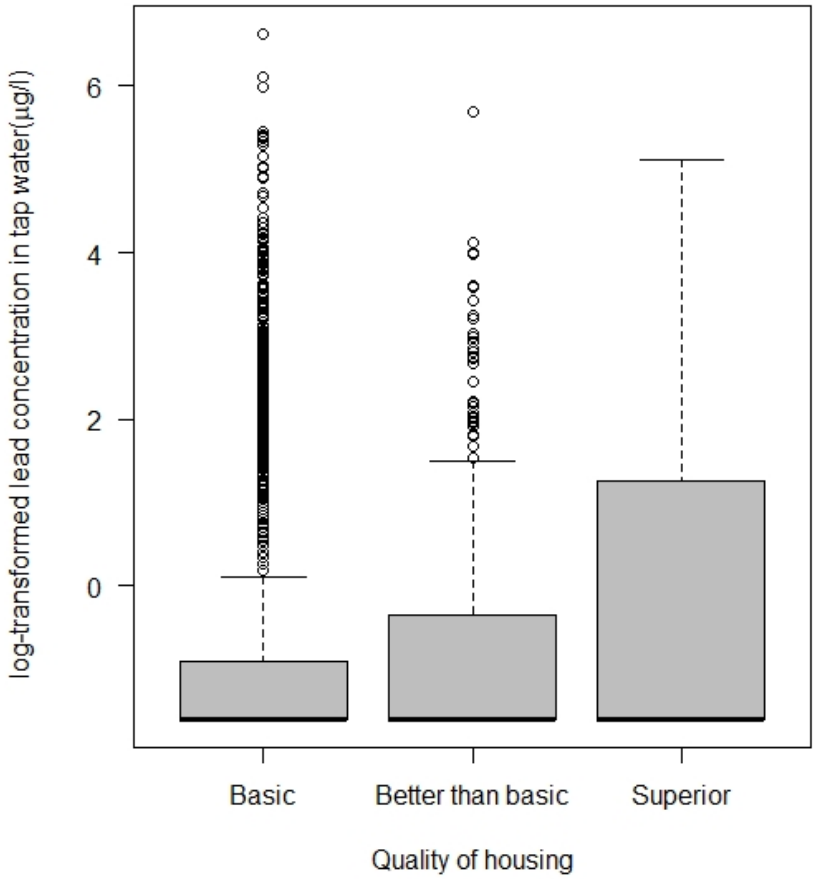

Figure 9: Boxplot showing relationship between house condition and lead concentration in tap water of houses surveyed within the same street postcode. Quality categories are defined by the SHCS where 'Basic' is the lowest level of house condition quality and 'Superior' is the highest. Resampled houses (n=3) and Grid Log survey data (n = 40) are excluded

Figure 9 shows the relationship between quality of houses surveyed by the SHCS and lead concentration in tap water of houses surveyed by Scottish Water within the same street postcode. There does not appear to be a significant difference in lead concentration in houses with different quality assessments (as defined by the SHCS). We used two separate models to examine how house condition related to lead values < 1 or $>=1$ μg/l and to model lead concentration values $>=1$ μg/ in relation to house condition. A binomial GLM showed that better quality houses were more likely to be found in street postcodes with houses with higher lead concentration values (Table 8), but there was no significant difference in lead concentration values of houses of different quality for lead concentration values $>=1$ μg/l (Table 9).

Table 8: Results of a binomial generalised linear model comparing lead concentration values (where values are <1 or >=1 μg/l) of houses with different quality assessments within the same street postcodes.

	Coefficient	Std error	z value	p-value
Intercept	-1.68	0.03	-52.78	<0.001
Better than basic	0.46	0.12	3.98	<0.001
Superior	1.52	0.32	4.72	<0.001

Table 9: Results of a gamma-distributed generalised linear model comparing lead concentration values (where values are >=1 μg/l) of houses with different quality assessments within the same street postcodes.

	Coefficient	Std error	z value	p-value
Intercept	2.47	0.11	21.58	<0.001
Better than basic	-0.16	0.40	-0.40	0.69
Superior	0.56	0.93	0.60	0.55

We also compared lead concentration values from different types of houses and flats using lead concentration data taken from houses surveyed in the same street postcode as SHCS respondents. Figure 10 shows that lead concentration appears to vary depending on type of house or flat.

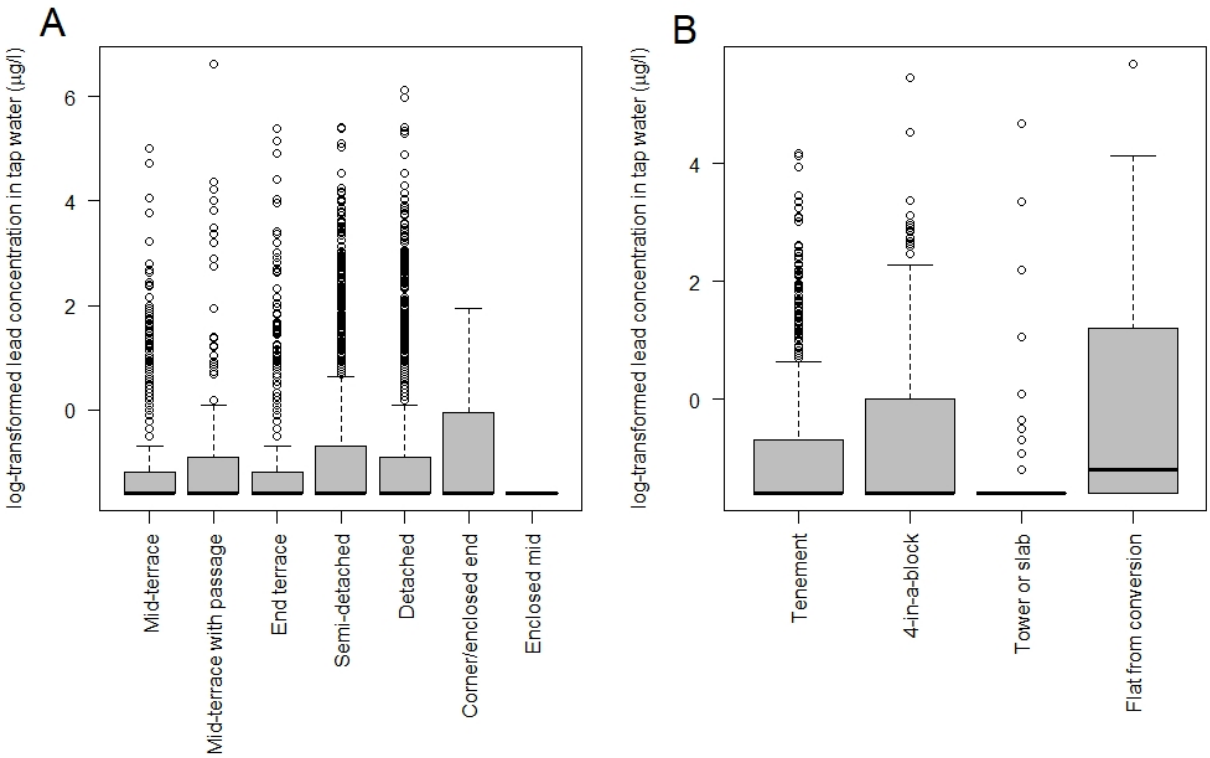

Figure 10: Plots exploring relationships between log-transformed lead concentration of tap water in houses and type of house/flat for A) Houses (Mid-terrace n = 659; Mid-terrace with passage n = 197; End terrace n = 584; Semi-detached n = 1754; Detached n = 3226; Corner/enclosed end n = 16; Enclosed mid n = 2); B) Flats (Tenement n = 808; 4-in-a-block n = 537; Tower or slab n = 74; Flat from conversion n = 110). Note that year house was built and lead concentration were not recorded for the same individual houses but for houses surveyed in the same street postcode. Resampled houses (n=3) and Grid Log survey data (n = 40) are excluded

We plotted log-transformed lead concentration of houses within street postcodes against orthophosphate dosing rate in the corresponding Water Treatment Zone. Figure 11 does not show an obvious relationship between lead concentration and orthophosphate dosing rate.

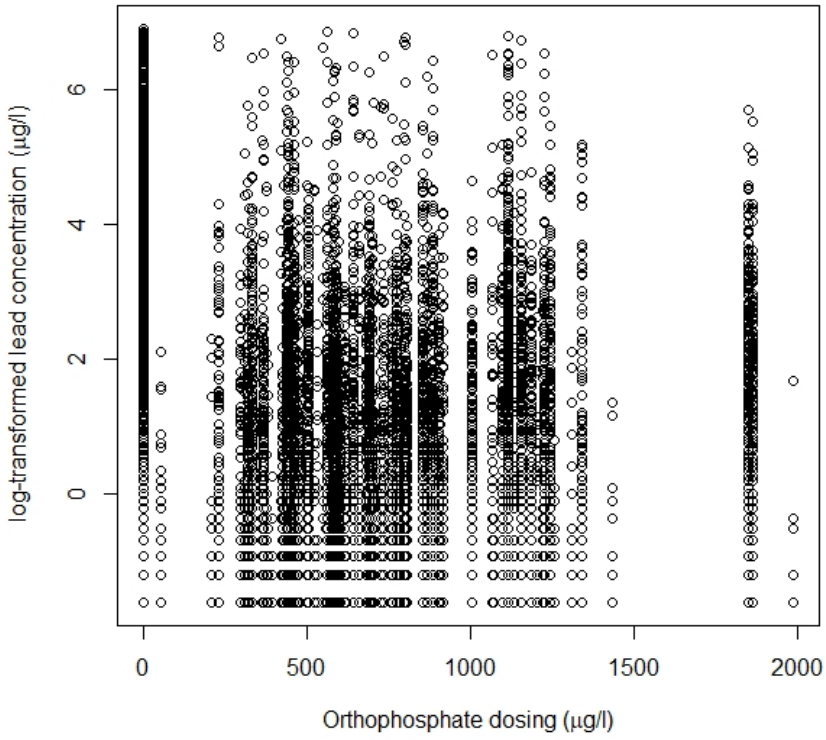

Figure 11: Plot showing relationship between log-transformed lead concentration of surveyed house and corresponding orthophosphate dosing rate at Water Treatment Zone level

Figure 12 shows the location of houses surveyed by Scottish Water for lead concentration of tap water from 2010 to 2018. We examined how many houses had been surveyed by Scottish Water for lead concentration of tap water in different district-level and street-level postcodes across Scotland. Figures 13 and 14 show the number of houses surveyed in different district and street postcodes respectively, while Figure 15 shows the overall distribution of these numbers at each spatial scale. Note that there are large numbers of street postcodes throughout Scotland for which no houses have been surveyed and that while multiple houses have been surveyed in some postcodes, in others only one or two houses have been surveyed. This makes aggregating lead concentration values over postcodes problematic as variation within postcode is likely to be under or over estimated in many cases.

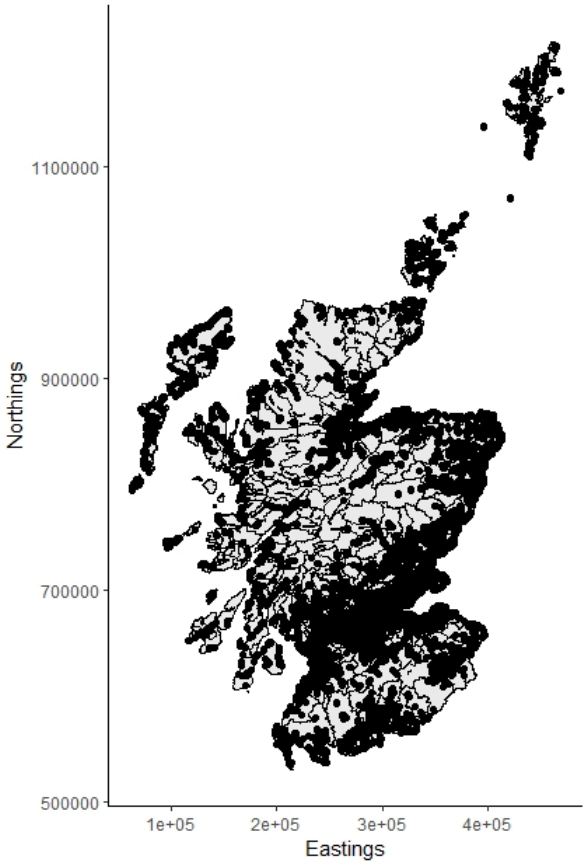

Figure 12: Location of houses surveyed for lead concentration of tap water from 2010 to 2018 per district postcode (in 2016)

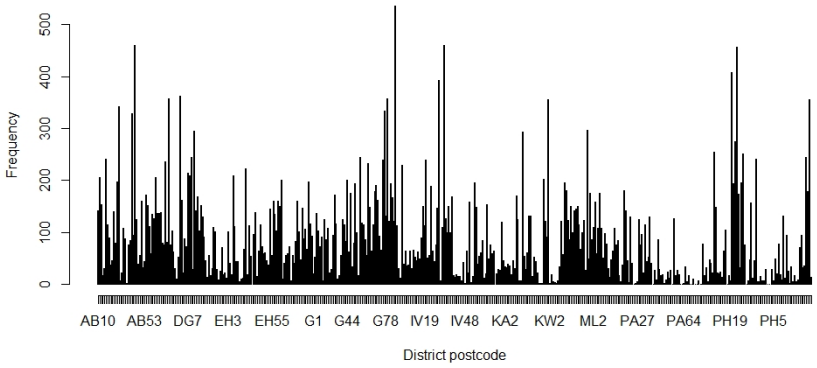

Figure 13: Number of houses surveyed for water quality by Scottish Water from 2010 to 2018 per district postcode

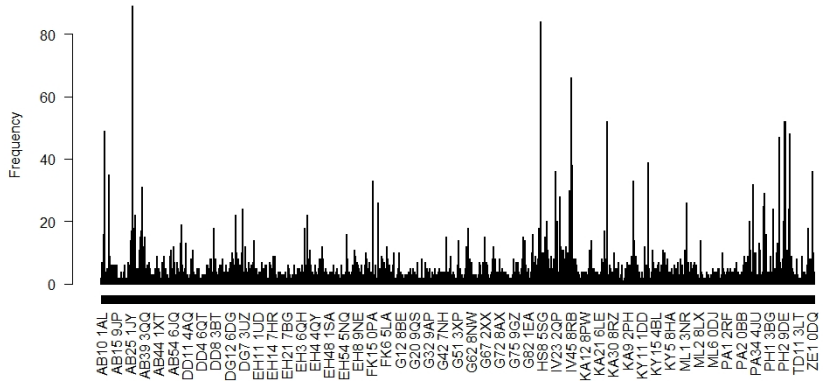

Figure 14: Number of houses surveyed for water quality by Scottish Water from 2010 to 2018 per street postcode

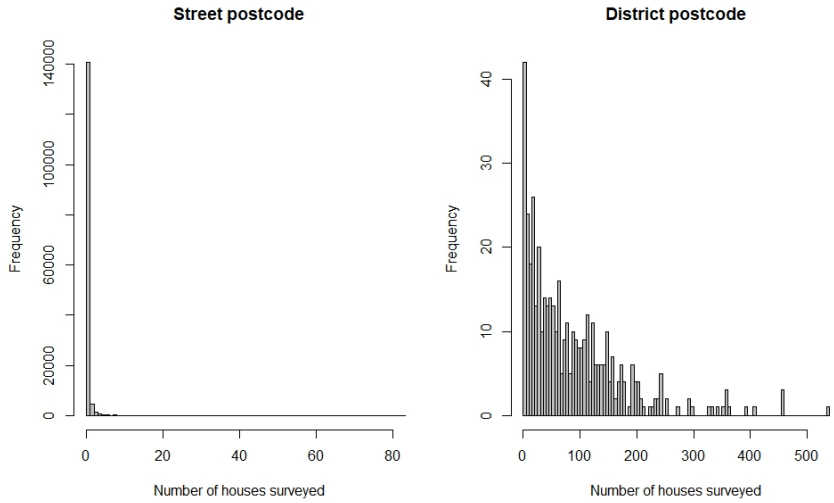

Figure 15: Histograms showing distribution of numbers of houses surveyed by Scottish Water per district and street postcodes from 2010 to 2018

Our exploratory analysis highlighted some issues/concerns that needed to be considered before formal analysis and modelling could begin:

- Our analysis shows that few respondents in the SHCS reported living in a house with lead piping. Relying solely on these data is likely to be result in an underestimate of the true numbers of houses with lead piping in Scotland, as few respondents are likely to know the lead status of their internal plumbing. Lead concentration of tap water data collected by Scottish Water may be a more appropriate response variable for use in formal models, although differences in survey methods used to collect these data should be accounted for. Bias may exist in response variables (presence of lead piping/lead concentration) due to sampling methods used to collect these data. To address this we used lead

concentration data that were collected using random surveys only ('Scheduled' samples). However, our exploratory analysis showed that Scheduled samples had higher lead concentration values than samples collected via Customer enquiries, which may mean that bias in the latter may be less than expected.

- Several houses have been surveyed (by Scottish Water and SHCS) in the same street postcode. We aggregated data per postcode (e.g. by calculating a mean or proportion) to avoid pseudoreplication in the dataset. Pseudoreplication artificially inflates the number of data points available for analysis and means that effects of interest cannot be statistically separated from effects due to variation among experimental units (in this case, street postcode). By aggregating data, variation within postcode is reduced allowing us to examine the effect of each variable on lead concentration. Our data exploration showed that many postcodes include less than five houses which were surveyed for lead concentration from 2010 to 2018, and fewer than these are likely to also include SHCS data. Aggregating these data will result in uncertainty being introduced to the model due to limited sample size per postcode. A possible solution to this problem would be to exclude from analysis any postcode for which fewer than five houses have been surveyed. The same also applies to houses surveyed via the SHCS. Although this reduces the size of the dataset to be analysed by creating a single entry per postcode and ignores variation within each postcode, this method reduces model complexity and allows results to be applied more easily to postcodes in which less data are available.

- In our exploratory analysis we included lead concentration data collected from 2010 to 2018. Including all years of surveys by Scottish Water from 2010 to 2018 does not correspond to survey years for SHCS (which are available for 2012 to 2017). However reducing the range of years for lead concentration data to 2012 to 2017 to match SHCS data results in significant reduction in the number of datapoints and number of unique postcodes containing lead concentration and SHCS data. Crucially, it reduces the number of postcodes which have both water quality data and houses reporting presence of lead piping from SHCS. For this reason, we used all water quality data available in further analyses.

- Lead concentration data have not been corrected for orthophosphate dosing in the above analysis. Although we did not find a significant relationship between amount of orthophosphate dosing at the water treatment plant and lead concentration in tap water, given our *apriori* knowledge of the effect of dosing on lead concentration we will include orthophosphate dosing levels in formal modelling procedures to control for any possible contributory effect on lead concentration in tap water.

- Note that the analyses described above are only preliminary analyses. Formal models accounting for biases in the data and including data over the correct temporal and spatial scales should be used before conclusions can be drawn.

6 Identifying subset of street postcodes for formal modelling

From recommendations made to us by the Drinking Water Quality Regulator (DWQR), we included water quality samples collected by Scottish Water from 2010 to 2018 and which were collected via Scheduled surveys only. By doing this we were able to include the maximum amount of data in our analysis, but by doing so we assume that explanatory variables such as house age and type remain consistent in postcodes between years. Our next aim was to select street postcodes for further analysis which included enough sampling points from which to aggregate data. Our data selection protocol following merging of Scottish Water and SHCS data is as follows:

- Check dataset and delete duplicate rows as far as possible
- Check number of points per postcode
- Select postcodes where sampled points comprise >=10% of all households in that postcode or >=7 total sampling points per postcode
- Map selected street postcodes
- Aggregate variables over each street postcode
- Carry out data exploration using aggregated data
- Build multivariable model explaining proportion of sampling points which tested positive for lead piping

6.1 Duplicates

The dataset to be used in formal analyses and modelling was created by merging separate datasets provided by Scottish Water and the SHCS. After including water quality data from Scheduled sampling only and excluding some variables which were not important for predicting lead concentration as identified in the exploratory analysis we had a dataset with 5861 rows and 406 unique postcodes. To reduce the number of duplicate rows introduced by merging datasets, we only included some variables from the communication pipe dataset which were similar for pipes within the same postcode (e.g. pipe age and material and whether or not the pipe had been replaced). Individual property information was not included (i.e. details of properties serviced by communication pipes). To ensure that there were an appropriate number of unique water quality sample points per street postcode we selected street postcodes from this dataset for inclusion in further analysis using the following criteria:

- Postcodes were selected where proportion of unique sampling points was >=0.1 of total households per postcode (based on 2011 Scotland census)
- Postcodes were selected where unique number of sampling points per postcode was >=7

These selection criteria resulted in a total of 100 selected street postcodes across Scotland.

6.2 Map selected postcodes

The location of all 406 street postcodes from the original dataset are shown in Figure 16. We mapped street postcodes selected as described above as shown in Figures 17. Note that a relatively large number of postcodes were selected from the outer Hebrides and Shetland. Before carrying out data exploration of this dataset we aggregated rows over each postcode to give a single variable value per street postcode (mean for a continuous variable and proportion per group for a categorical variable). As explained above, we aggregated data in each postcode (e.g. by calculating a mean or proportion depending on whether data were continuous or categorical) to avoid the problem of pseudoreplication in the dataset and allow models to be more easily applied to future datasets.

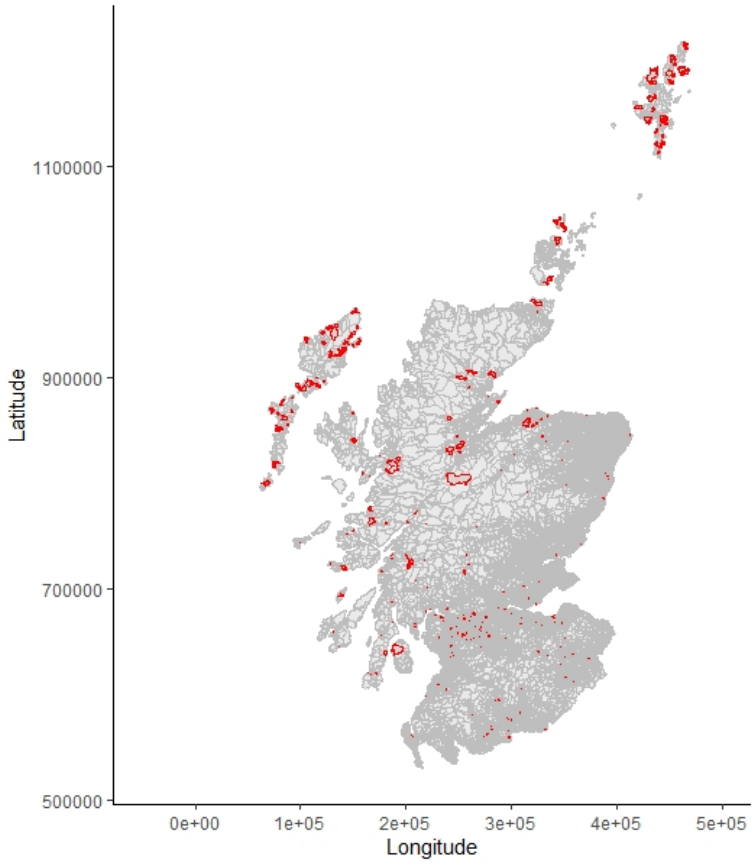

Figure 16: Map showing location of all 406 street postcodes

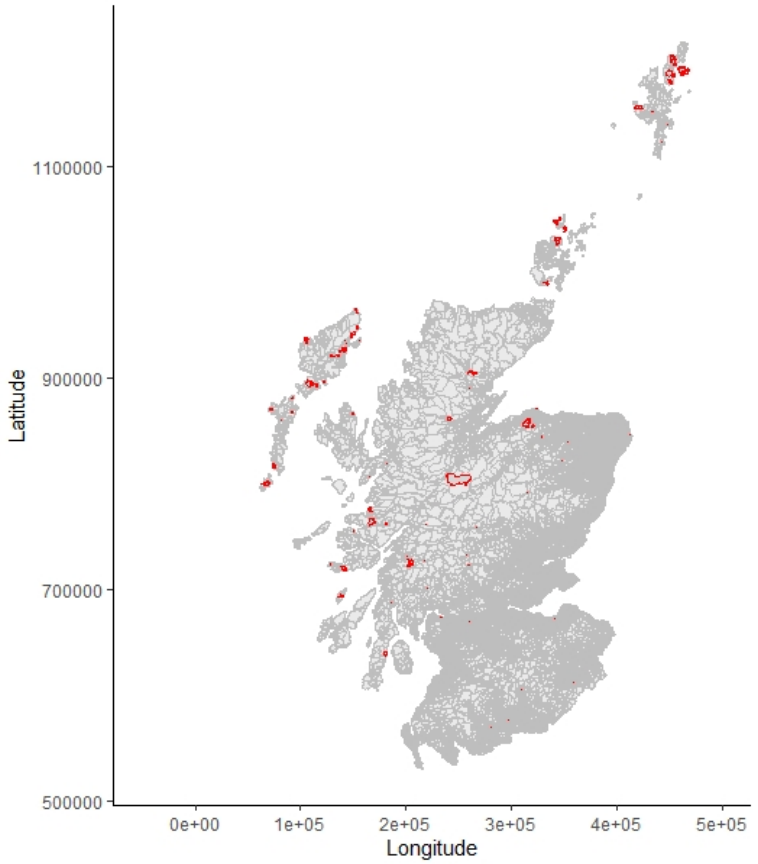

Figure 17: Map showing location of 100 street postcodes selected based on the criteria described above (proportion of sampling points were >=0.1 of total households per postcode or had >=7 sampling points in total)

6.3 Data exploration for aggregated data

First we explored the relationship between mean house age per street postcode and number of sample points which tested positive for lead per postcode (according to a threshold where lead concentration values >= 1 μg/l were regraded as positive and values < 1 μg/l were regarded as negative for lead piping) out of total number of sample points. Figure 18 shows a noisy relationship between proportion of positive sampling points and mean house age per postcode. There may be a positive relationship between the two variables, but the large number of zero proportions could make this relationship difficult to detect. The distribution of proportions of samples positive for lead per postcode is shown in Figure 19.

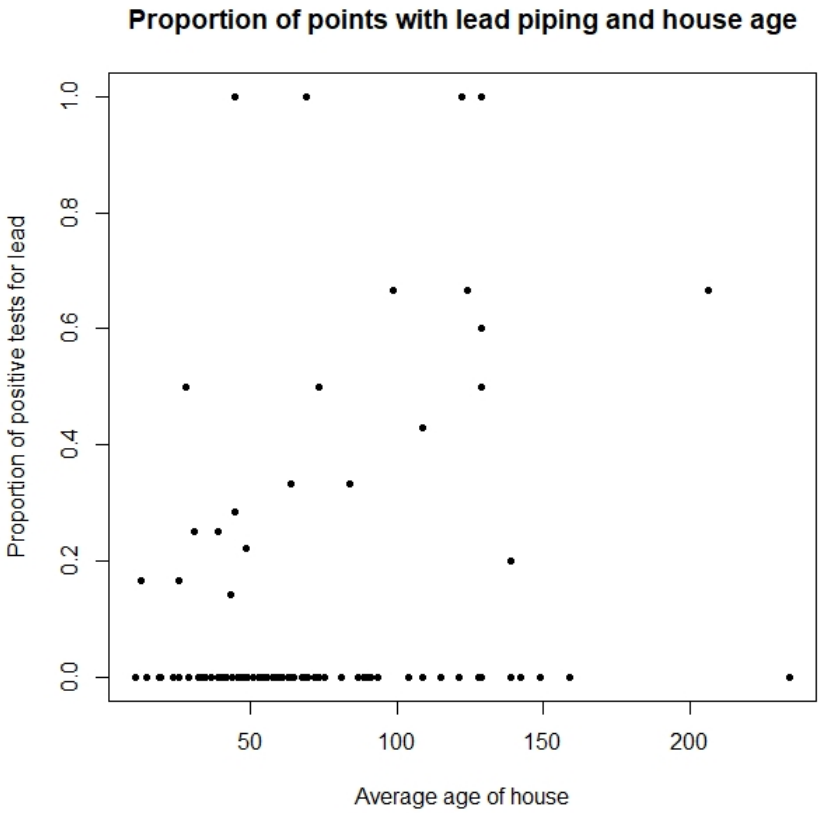

Figure 18: Plot showing relationship between proportion of sampling points which tested positive for lead and mean house age per street postcode (years) (n = 100)

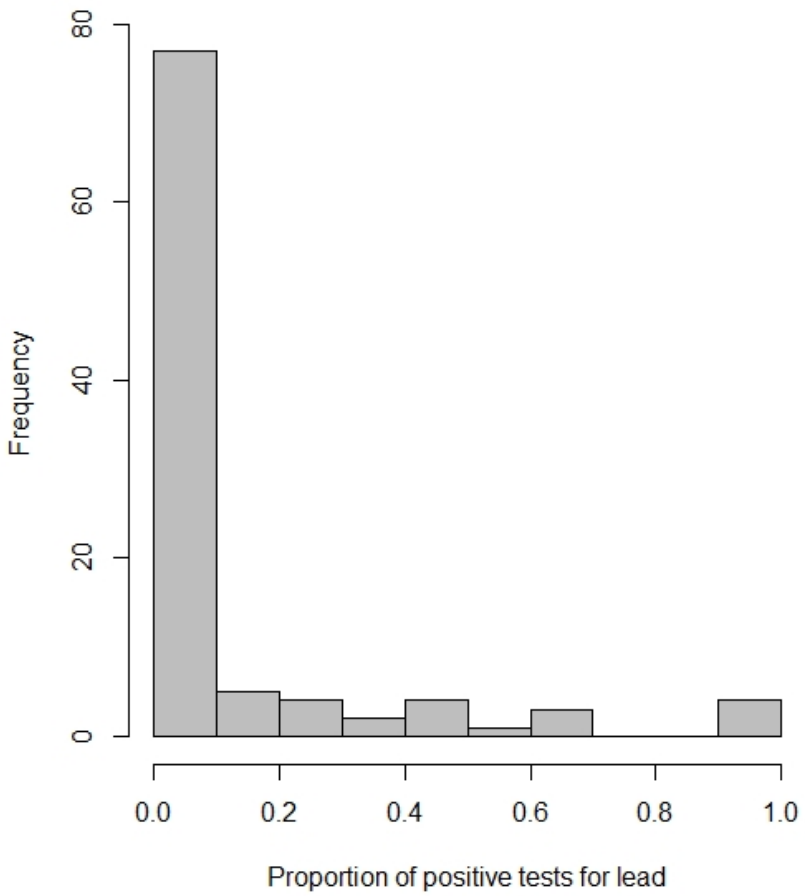

Figure 19: Histogram of proportion of sampling points which tested positive for lead per street postcode (n = 100)

Due to the large number of zero proportions in the data (referring to postcodes which have no houses which tested positive for lead according to our threshold value), we used a zero-inflated negative binomial model to examine the relationship between number of sample points which tested positive for lead per postcode and mean house age, accounting for the total number of sample points per postcode. The results of the model are shown in Table 10. The model suggests that house age is an important variable in explaining variation in number of positive samples for lead piping and should be included in the final model.

Figure 20 shows the relationship between proportion of sample points with lead and proportion of communication pipes made of lead and average age of pipe per street postcode. There do not appear to be strong relationships involving communication pipe variables. A zero-inflated negative binomial model showed neither variable to be significant. We also examined the relationship between proportion of sample points with lead and mean orthophosphate dosing per street postcode (see Figure 21), although there is no clear

relationship between these two variables. Despite this, level of orthophosphate dosing was included in the multivariable model as a potential confounding variable.

Table 10: Results of a zero-inflated negative binomial model comparing number of sample points positive for lead per postcode (where values are >=1 μg/l) offset for total number of points sampled per postcode and mean house age per postcode.

Count model	Coefficient	Std error	z value	p-value
Intercept	-1.79	0.26	-7.03	<0.001
House age	0.008	0.003	2.87	0.004
Binomial model	Coefficient	Std error	z value	p-value
Intercept	0.64	0.55	1.15	0.25
House age	0.001	0.006	0.20	0.84

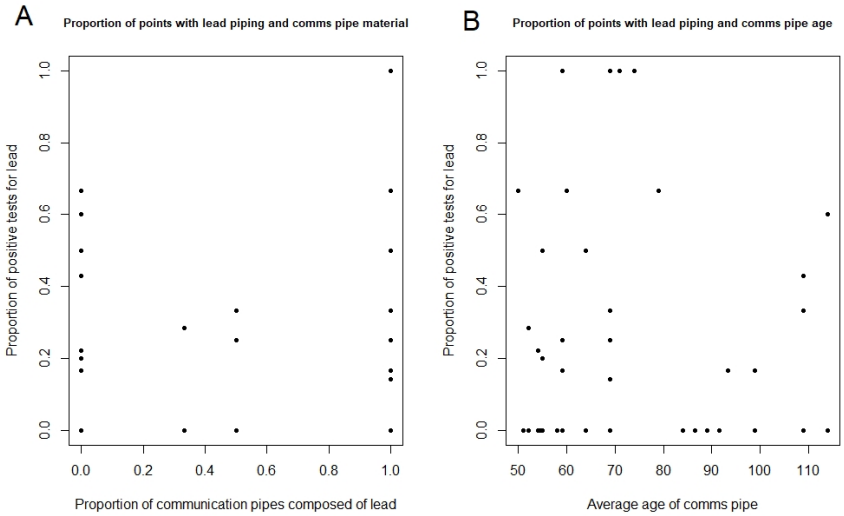

Figure 20: Plot of proportion of sampling points which tested positive for lead and A) proportion of commincation pipes composed of lead and B) mean age of communication pipes per street postcode (n = 100)

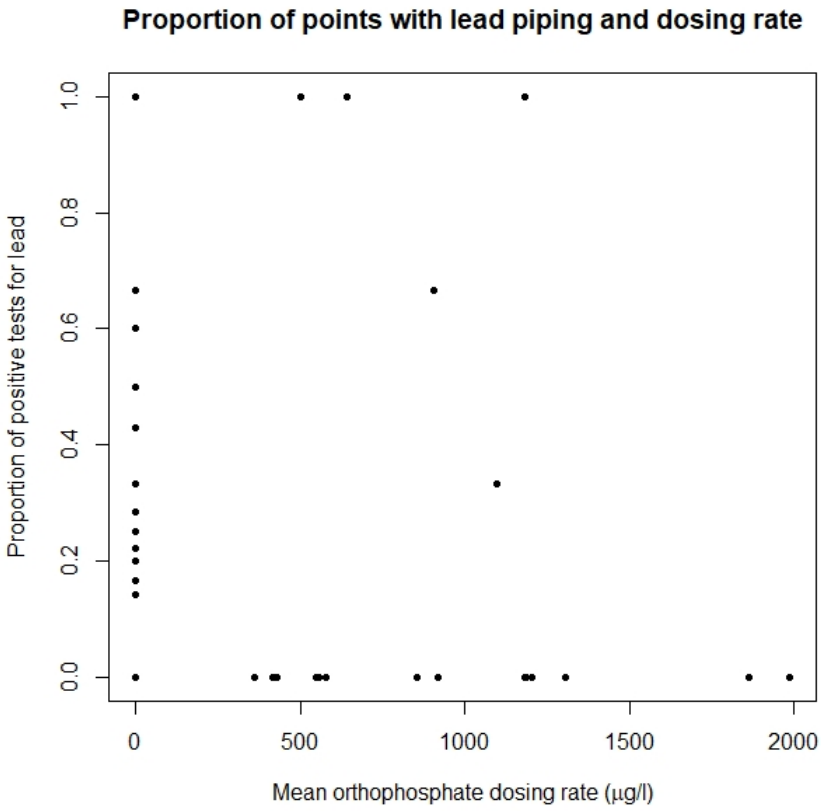

Figure 21: Plot of proportion of sampling points which tested positive for lead and mean orthophosphate dosing rate per street postcode (n = 100)

7 Using the new house age dataset

7.1 House age per postcode

The Scottish Assessors Association (SAA) made a dataset available which has house ages for houses randomly sampled in each street postcode. The number of houses sampled per postcode varies from 1 to over 10. This dataset covers a larger number of postcodes in Scotland than the SHCS dataset, although no house age data was available from the SAA in the Highlands and Western Isles. We decided to use this data source in further analysis as more house age data are available from more postcodes throughout Scotland than from the SHCS.

7.2 Formal modelling

We used the zero-inflated model described above to initially identify important explanatory variables which were then selected for use in the formal predictive model. We repeated the postcode selection process for the formal model using the dataset provided by the SAA for house age data. We merged datasets which

contained important variables identified during exploratory data analysis described above. The important variables identified by the zero-inflated model were house age, whether or not a postcode contained a communication pipe which had been replaced by Scottish Water, orthophosphate dosing level and proportion of houses in a postcode which were built before 1970. Datasets containing these variables were merged together to give a dataset with 96,316 rows and 5007 unique postcodes. We selected postcodes from this dataset where proportion of sampling points was >=0.1 of total households per postcode (based on 2011 Scotland census) or where number of sampling points per postcode was >=7 resulting in a dataset with 29,500 rows and 1347 unique street postcodes (see Figure 22). From this dataset we selected only street postcodes that had >=10 houses with known ages (from SAA data) (see Figure 23). This final selection resulted in a dataset with 21,600 rows and 312 unique street postcodes. No postcodes in the Highlands or Western Isles were selected as there are no SAA house age data for these regions. Finally, we included number of sampling points per postcode that were positive for lead and total number of points sampled per postcode in this dataset. We excluded four street postcodes which included multiple samples taken from the same house, leaving a dataset with data from 308 postcodes which was used in formal modelling (see Figure 24).

From this dataset we aggregated each variable by street postcode calculating means and medians for continuous variables (house age, orthophosphate dosing rate) and proportions for categorical variables (age category (pre-1970, post-1970), communication pipe replacement).

We explored relationships between these variables (see Figures 25 and 26) before including them together in a hurdle model with a Poisson distribution, to account for zero-inflation in the response variable. We chose a hurdle model for this analysis to account for the large number of postcodes in which no houses had lead piping (according to our threshold of 1 μg/l). The hurdle model assumes zero values are all non-structural which coincides with what we believe about the proportion of houses with lead pipes per postcode. We modelled number of sample points per postcode that tested positive for lead (i.e. was over the threshold of 1 μg/l) accounting for total number of samples taken per postcode, including previously identified important explanatory variables, such as median house age as estimated from SAA data. Total number of samples collected per postcode was included as an offset in the model and the response variable was number of sample points per postcode which tested positive for lead. Explanatory variables included were median house age, proportion of houses built pre-1970, orthophosphate dosing rate and proportion of communication pipes replaced in that postcode.

We compared Poisson-distributed and negative binomial-distributed hurdle models using Akaike's Information Criterion (AIC). We found that the Poisson distributed hurdle model had a lower AIC (>2 points) than the negative binomial hurdle model and overdispersion was not found to be affecting results, hence the former was selected for further analysis. We used an ANOVA-based model selection procedure to assess the importance of each individual explanatory variable in explaining variation in number of samples which tested positive for lead. We found that proportion of houses built pre-1970 and proportion of communication pipes replaced per postcode were significant and these were included in the final predictive model. Hurdle models including different variables in binomial and truncated Poisson sections were compared using AIC. The model with the lowest AIC included all explanatory variables in both the binomial and Poisson-truncated sections

of the model and a Poisson-distributed hurdle model including all explanatory variables in both sections of the model was used in further analysis.

To formalise the Poisson-distributed hurdle model used, let $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{N}^d$ be a random vector where each Y_i represents the total number of households in the ith postcode in a given period of time. Alongside $\boldsymbol{Y}$, a matrix of covariates $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n)^\top \in \mathbb{R}^{n \times k}$ is also given. We facilitate modelling of covariates via a generalized linear model (GLM) with Poisson hurdle probability mass function. The model is based on the assumption that when all covariates are held fixed, the generation mechanism for zero counts is a Bernoulli trial with covariate-dependent probability of success, whilst the generation mechanism for positive counts follow a zero-truncated Poisson distribution with covariate-dependent rate parameter. Writing $\pi(x) = \{1 + \exp(-x)\}^{-1}$ and $\lambda(x) = \exp(x)$, $x \in \mathbb{R}$, for the canonical link functions of the logistic and log-linear part of the model respectively, the Poisson hurdle GLM is specified according to its conditional probability mass function as

$$\Pr(Y_i = y \mid \boldsymbol{Z}_i = \boldsymbol{z}) = \begin{cases} 1 - \pi(\boldsymbol{z}^\top \boldsymbol{\beta}_0) & y = 0 \\ \pi(\boldsymbol{z}^\top \boldsymbol{\beta}_0) \dfrac{\exp\{-\lambda(\boldsymbol{z}^\top \boldsymbol{\beta}_+)\} \lambda(\boldsymbol{z}^\top \boldsymbol{\beta}_+)^y}{y! \, [1 - \exp\{-\lambda(\boldsymbol{z}^\top \boldsymbol{\beta}_+)\}]} & y \in \mathbb{N}_{\setminus 0}, \end{cases} \tag{1}$$

where $\mathbb{N}_{\setminus 0} = \{1, 2, \ldots\}$, and $\boldsymbol{\beta}_0 = (\beta_{00}, \ldots, \beta_{0k})^\top \in \mathbb{R}^{k+1}$ and $\boldsymbol{\beta}_+ = (\beta_{+0}, \ldots, \beta_{+k})^\top \in \mathbb{R}^{k+1}$ are vectors of parameters describing the effect of covariates for the zero and hurdle part of the model respectively. The maximum likelihood estimator $(\hat{\boldsymbol{\beta}}_0, \hat{\boldsymbol{\beta}}_+)$ is obtained by solving a system of $2(k+1)$ equations. Since there is no analytic form for the estimator, numerical methods such as the BFGS iterative method, are typically used to obtain values for $\hat{\boldsymbol{\theta}}$ [1]. In a Bayesian setting, inference proceeds by specifying a prior distribution over $(\boldsymbol{\beta}_0, \boldsymbol{\beta}_+)$.

Of the variables included in the model, only the proportion of houses built pre-1970 and the proportion of replaced communication pipes per postcode were significant. Model diagnostic plots suggest that the model fitted the data well (see Figure 27). Model outputs displayed in Table 11 suggest that greater numbers of positive lead sampling points are found in postcodes with higher proportions of pre-1970 houses and replaced communication pipes.

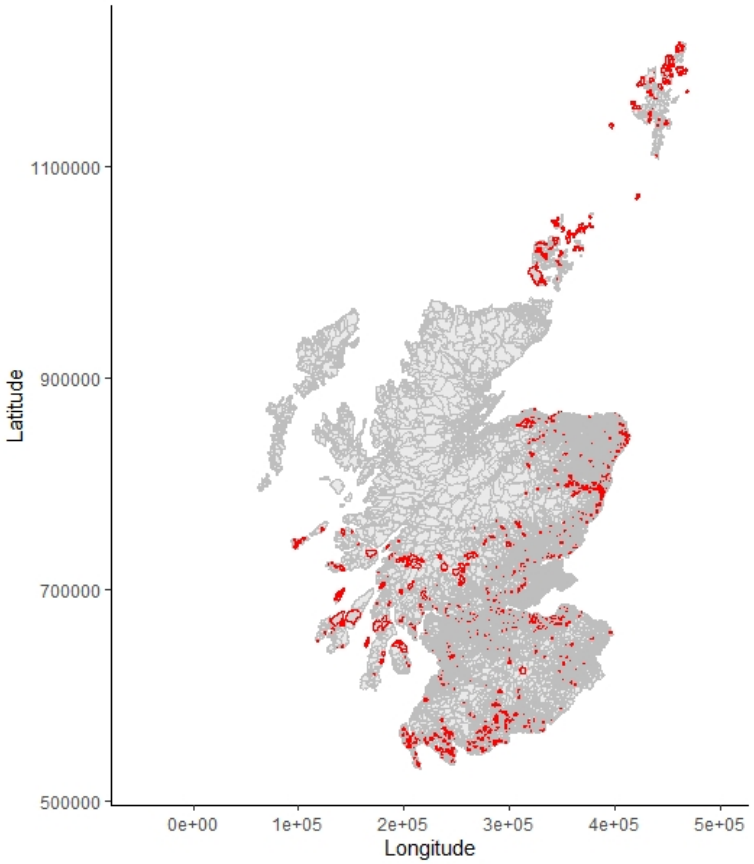

Figure 22: Map showing locations of all 1347 street postcodes selected based on stated criteria (proportion of sampling points were >=0.1 of total households per postcode or had >=7 sampling points in total) (excluding those with no data on number of dwellings (n=9))

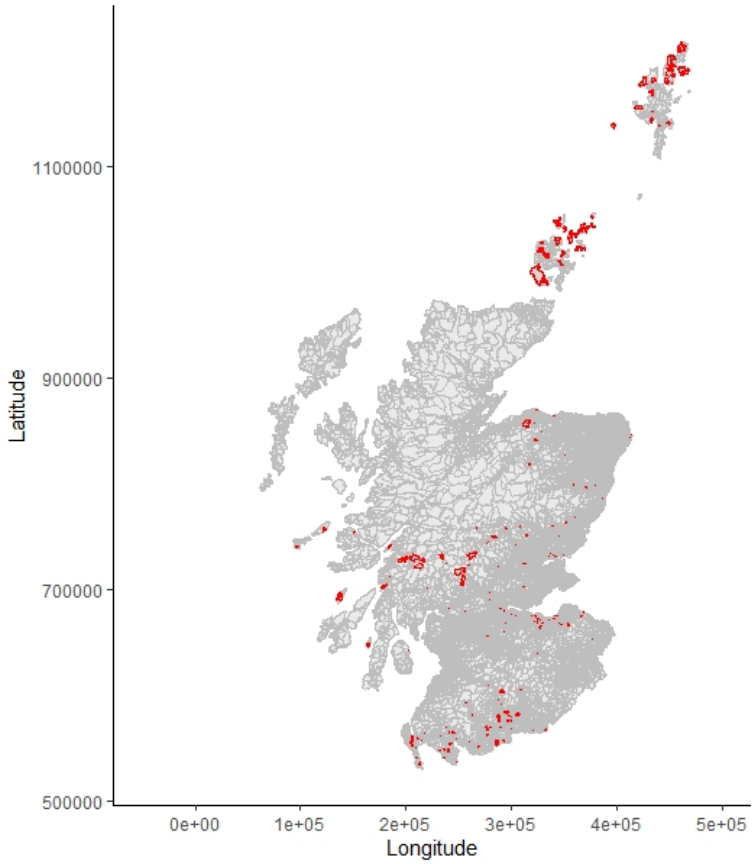

Figure 23: Map showing locations of selected street postcodes with SAA house age data for >=10 houses per postcode

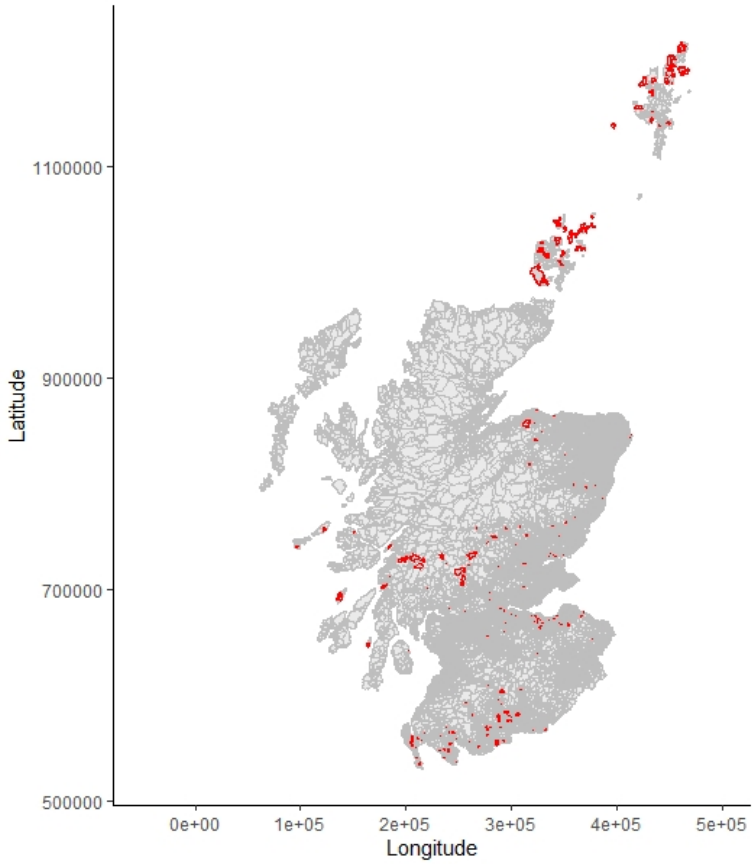

Figure 24: Map showing locations of 308 selected street postcodes with SAA house age data for >=10 houses per postcode and excluding postcodes with samples taken from the same house

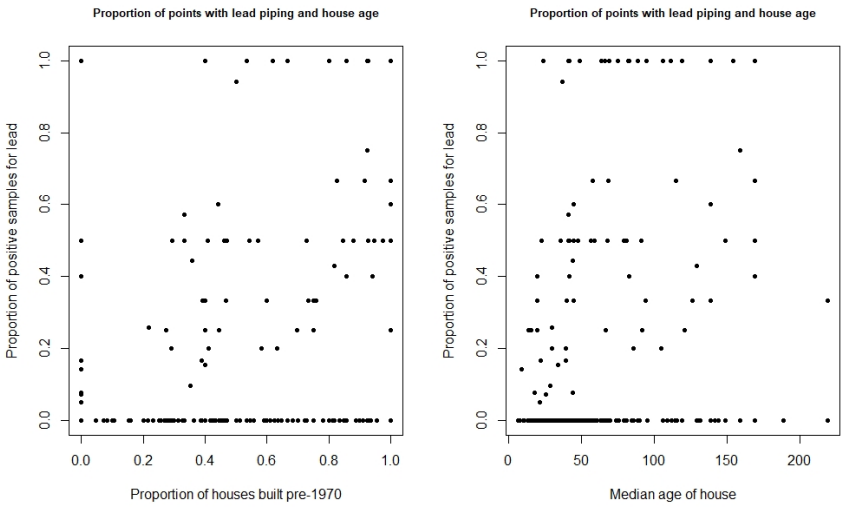

Figure 25: Plot of proportion of sampling points which tested positive for lead and A) proportion of houses per postcode built pre-1970 and B) median age of houses per street postcode (n = 308)

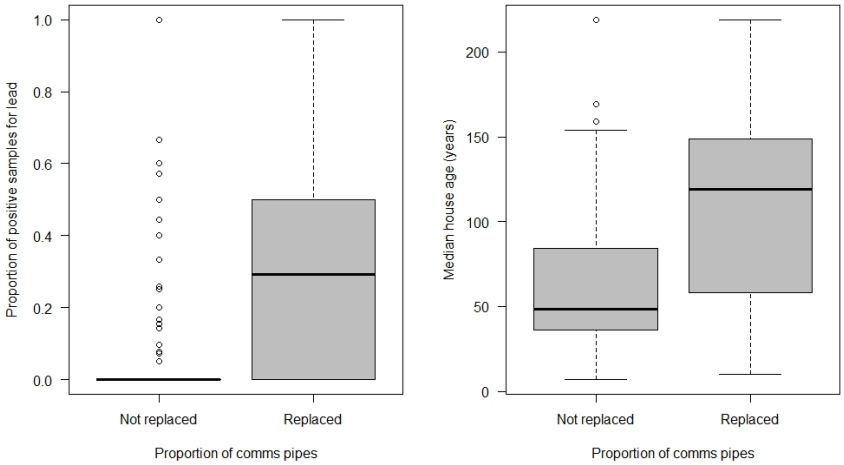

Figure 26: Plots of A) proportion of sampling points which tested positive for lead and proportion of postcodes in which communication pipes have been replaced and B) median age of houses per street postcode and proportion of postcodes in which communication pipes have been replaced (n = 308)

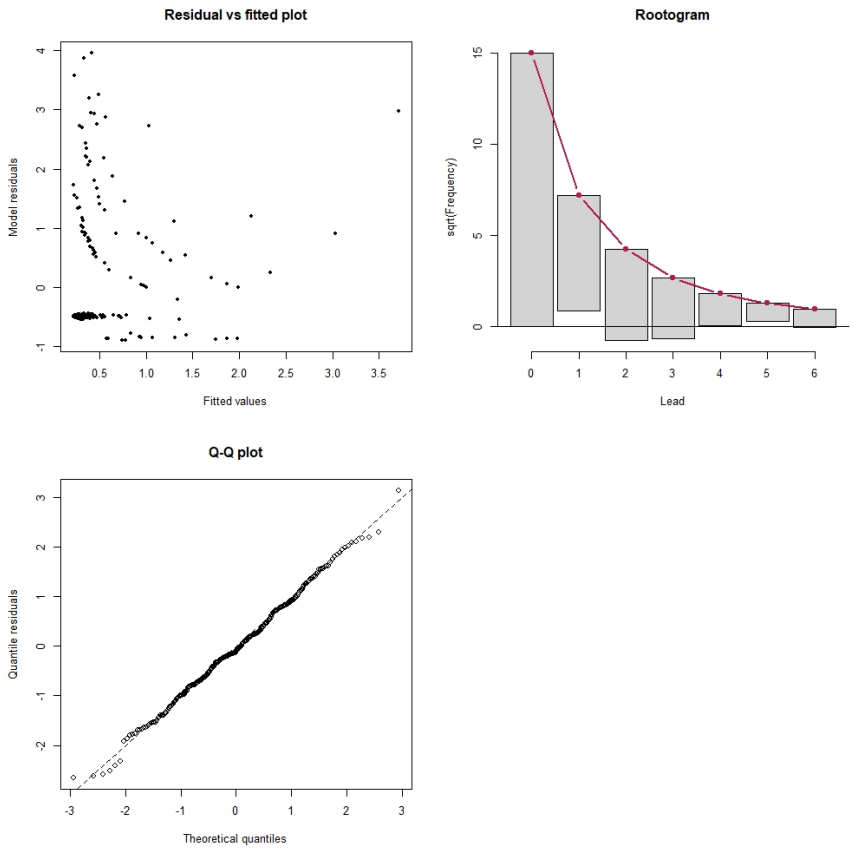

Figure 27: Diagnostic plots of final model showing A) Residual vs fitted values B) Rootogram goodness-of-fit plot C) Q-Q plot of model residuals (n = 308)

Table 11: Results of Poisson-distributed hurdle model comparing number of sample points positive for lead per postcode (where positive values are >=1 μg/l) offset for total number of points sampled per postcode. The binomial section of the model shows the effect each explanatory variable has on whether or not a given postcode contains houses with lead concentration thresholds >=1 μg/l and the truncated-Poisson section shows the effect of explanatory variables on number of sample points positive for lead per postcode (where values are >=1 μg/l). The final model following model selection includes proportion of pre-1970 houses per postcode and proportion of communication pipes replaced per postcode as explanatory variables. N = 308

Count model	Coefficient	Std error	z value	p-value
Intercept	-2.00	0.22	-9.11	<0.001
Pre-1970	1.13	0.35	3.24	0.001
Replaced	0.59	0.22	2.65	0.008
Binomial model	Coefficient	Std error	z value	p-value
Intercept	-1.35	0.24	-5.64	<0.001
Pre-1970	0.34	0.36	0.95	0.34
Replaced	1.20	0.38	3.14	0.002

8 Applying the model to street postcodes across Scotland

We used the output from the hurdle model described above to predict the mean number of houses with lead piping (i.e. with tap water samples >=1 μg/) per street postcode throughout Scotland. The SAA dataset provided data on the year selected houses per postcode were built and age category to which each house belonged. Where data on what year a house was built were not available we used age category to identify houses built pre- or post-1970. Data on house age (from SAA dataset) and whether or not communication pipes have been replaced in each postcode (from Scottish Water datasets) were available for 76,546 postcodes across Scotland. Of these, 70,180 had age or age category data available for >=10% of houses per postcode or >=7 houses with age data per postcode. We made predictions of mean numbers of houses with lead piping for each group of postcodes (those with any houses per postcode with age data (76,546) and those including >=10% of aged houses (70,180)). For each mean prediction per postcode we estimated the likely maximum number of houses with lead piping per postcode by calculating 95% prediction intervals within which 95% of future predictions may be expected to fall (from zero houses with lead to maximum number of houses with lead per postcode).

We mapped model predictions and upper 95% prediction intervals for both groups of postcodes as shown in Figure 28. Age data were missing for many postcodes across Scotland according to street postcode definitions as of 2019 and postcodes with missing data are shown in green on maps.

We have model predictions for 40% or 37% of total street postcodes in Scotland depending on whether we included all available postcodes or postcodes selected using the criteria outlined above. Depending on which groups of postcodes were included in the analysis, our model predicts that totals of 97,975 or 109,500 houses in Scotland have lead pipes (using all or selected postcodes respectively). These values are calculated by summing up predicted mean number of houses with lead piping in each street postcode across the country. Our model also provides an estimate of the maximum numbers of houses in Scotland with lead pipes according to 95% prediction intervals calculated from the model for each postcode. The output given by the model is a mean predicted number of houses with lead piping per street postcode (i.e. with lead concentrations >1 μg/) and an estimated maximum number of houses with lead piping per postcode. We can expect with 95% confidence that any new estimate of the number of houses with lead piping will be below this maximum should new data be collected and the study repeated. Our model output estimates the total number of houses expected to have lead piping according to model predictions. Depending on whether we included data from all available postcodes or only those selected according to previously stated criteria (see above), maximum numbers of houses with lead pipes according to 95% prediction intervals were 396,213 and 470,726 respectively.

As stated above model predictions were made for only 37% to 40% of total street postcodes in Scotland (of which there were 190,700 in 2019). To estimate the total number of houses in Scotland with lead piping we extrapolated our estimates of the number of houses with lead piping in 37% or 40% postcodes to all street postcodes in Scotland to give a mean of 273,751 and 264,532 houses with lead pipes according to estimates from all available postcodes (76,546) and selected postcodes (70,180) respectively (with corresponding maximum estimates according to 95% prediction intervals of 1,397,912 and 1,232,554). We assume that

the remaining 60% and 63% of postcodes which had no model predictions had a similar distribution of prediction values to those made in postcodes with data available.

Making predictions for all postcodes with explanatory variable data may lead to some inaccurate estimates if only a small proportion of houses in the postcode had age data. Only making predictions of number of houses with lead piping in postcodes in which more than 10% of houses had an age estimate leads to more accurate estimates of numbers of houses with lead pipes. Hence our estimate of 264,532 houses with lead pipes in Scotland may be expected to be more accurate (with upper range of 1,232,554 according to 95% prediction intervals).

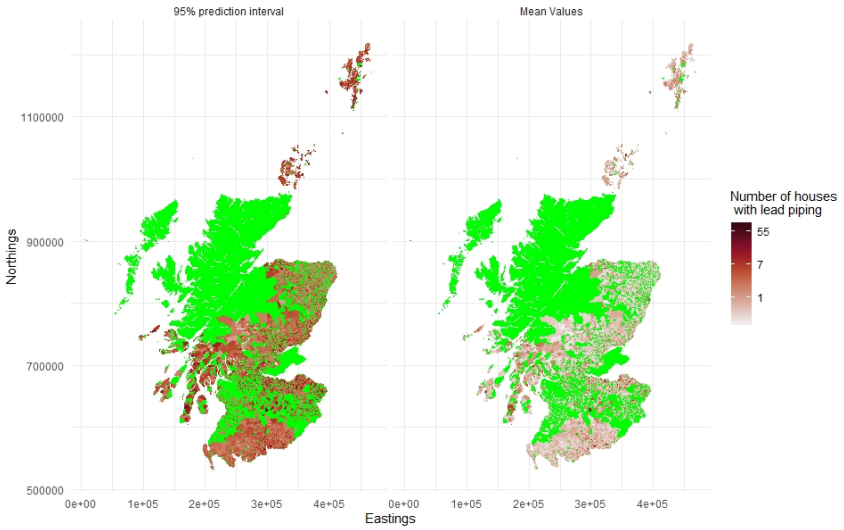

Figure 28: Map showing predicted numbers of houses with >=1 μg/l for 76,546 street postcodes including house age data. Green represents areas with no explanatory variable data

9 Sampling methodology

The model predicts the mean number of houses in each street postcode with >=1 μg/l of lead in tap water and uncertainty around these predictions. We aimed to validate the predictions made by our model and determine how well our model fits actual data in specific regions of interest (namely Glasgow) by taking tap water samples from postcodes throughout Scotland. Due to logistical difficulties in retrieving samples from islands, postcodes located in Shetland, Orkney, the Western Isles and Arran were excluded from sampling. Our sampling protocol ensured that sample kits were sent to addresses in postcodes which were included in the model fitting process as well as postcodes which were not included in the fitting process but to which the model was applied. This should enable us to assess how well the model was fitted and the validity of subsequent extrapolations. Postcodes with both high and low percentages of houses with lead pipes were equally represented in the samples. The first round of sampling was carried out from January to March 2020 for which 300 street postcodes were selected for sampling.
In total, the model was fitted using 308 postcodes, 227 of which were not located on islands. A total of

68,546 non-island postcodes had predictions of mean number of houses with lead extrapolated from model output.
The sampling method used is described as follows:

- 49 postcodes (from non-modelled postcodes) were randomly selected from the Glasgow region. These were included together with a single postcode from which the model was fitted in Glasgow. Random selection permitted re-sampling allowing the possibility that the same postcode may be selected for sampling more than once (in which case sampling kits would be sent to more than one address in that postcode).

- 100 postcodes from which the model was fitted were randomly selected without allowing for re-sampling in other areas of Scotland (excluding islands). Of these postcodes, we used stratified sampling to ensure that postcodes with high and low percentages of lead houses were equally represented (50 postcodes where >=10% of the houses were predicted to have lead and 50 postcodes where <10% of postcodes were predicted to have lead). We excluded postcodes with less than 10 houses in total (according to 2011 census); this ensured that enough houses were available for sampling if the house selected did not provide a sample and that the calculation of percentage of houses with lead per postcode was less prone to bias from small sample sizes.

- As there are more postcodes which were not used to fit the model than postcodes which were used to fit the model, 150 non-modelled postcodes were randomly selected, where half were selected from postcodes with higher percentages of lead houses (>=10%) and the other half from postcodes with lower percentages of lead houses (<10%).

We checked that the 300 postcodes selected were still in use using the Royal Mail's 'Postcode and Address Finder' (`https://www.royalmail.com/business/find-a-postcode`), the 'click to address' website (`https://craftyclicks.co.uk/address-finder-from-postcode/`) and Zoopla's address finding website (`https://www.zoopla.co.uk/postcode-finder/`). One address was selected for sampling from each postcode using address finding websites.

The maps in Figure 29 show the locations of the 300 postcodes selected for tap water sampling.

Figure 29: Maps showing locations of 300 postcodes randomly selected for the first round of sampling. One house per selected postcode had a sample of tap water taken to be tested for lead

From this sampling methodology, we expected to find an overall mean of 30 houses with lead values >=1 μg/l in our sample of 300 houses (assuming that samples are returned from every postcode) as a mean of approximately 10% of houses had lead pipes per postcode according to our model both for postcodes across Scotland and for 300 sample postcodes. This mean estimate has a standard deviation of 27 houses (between 3 and 57 houses are expected to have lead concentration values >=1 μg/l).

10 Collecting samples

Three hundred sample kits and information packs were sent out to addresses within our 300 selected postcodes (to one address per postcode). Each kit included a plastic resealable bottle to collect water samples in, a stamped addressed envelope for returning the sample to the lab at the University of Edinburgh and a detailed letter explaining the purpose of the study, the reasons for selecting this address for sampling, instructions for taking a water sample and how to return the sample after it has been taken. Instructions in the letter included photographic instructions to make the tap water sampling procedure as clear as possible to all respondents, some of whom may experience language or reading barriers. Contact details for the project

manager at the University of Edinburgh and Scottish Water were included in the letter so that participants were able to contact us for more information about the project. We intended the sampling process to be as simple and easy to follow as possible to maximise the number of respondents and samples returned. Despite this, based on previous studies we only expected to receive between 30% and 40% of sample kits that were given out. We did not experience many problems with returned samples, but on some occasions respondents had not screwed the top on the sample bottle correctly and some water had leaked out during transport. As we asked for a large water sample we still had a large enough water to successfully run lab analyses on these samples.

11 Water sample analysis

12 Model validation

Using the sampling regime described above, we received tap water samples from 155 unique postcodes out of 300 postcodes to which water sampling kits were sent. Chemical analysis was carried out on 205 water samples in total (some samples were tested more than once to ensure the consistency and accuracy of the results as the analyses progressed). Where a water sample from a single address within a single postcode was tested more than once we calculated the mean concentration of lead over all samples tested from that address. This gave a total of 155 sample results some of which were means calculated from more than one analysis of a water sample. From this we calculated that approximately 15.5% of postcodes had samples that had a lead concentration of >=1 μg/l (24 out of 155 postcode samples). Of postcodes from which samples were returned, a mean of 10.7% (with standard deviation of 10.8%) positive samples were expected to be returned according to model predictions.

Water samples were returned from 69 of 125 sampled postcodes in which >=10% of houses were estimated to have lead piping (lead concentration of tap water >=1 μg/l) and 20.3% (n = 14) of samples from these postcodes had lead concentrations >=1 μg/l. Using model predictions only from postcodes for which samples were returned, we estimate that a mean of 18.3% of houses should return positive samples (with a standard deviation of 12%). Water samples were returned from 64 of 125 postcodes in which <10% of houses were predicted to have lead piping and 10.9% (n = 7) of these samples had lead concentrations of >=1 μg/l. Of postcodes with an estimated lower percentage of houses with lead pipes for which samples were returned we estimated that a mean of 4.5% (with standard deviation of 2%) of houses within each postcode had lead pipes. Twenty-two samples were returned from 50 postcodes selected from Glasgow and 13.6% (n = 3) of these had lead concentrations >=1 μg/l compared to a mean of 5% of houses within these postcodes (with standard deviation of 6.1%) according to model predictions.

Figure 30 shows lead concentration values from water samples returned from postcodes with predicted high and low percentages of houses with lead piping and from postcodes samples in Glasgow. Figure 30 suggests that water samples returned from postcodes estimated to have a high percentage of houses with lead piping had higher lead concentration values than postcodes with a low percentage of houses with lead piping, and this difference was found to be significant (Mann-Whitney U test: U = 1708, p = 0.02, n = 133).

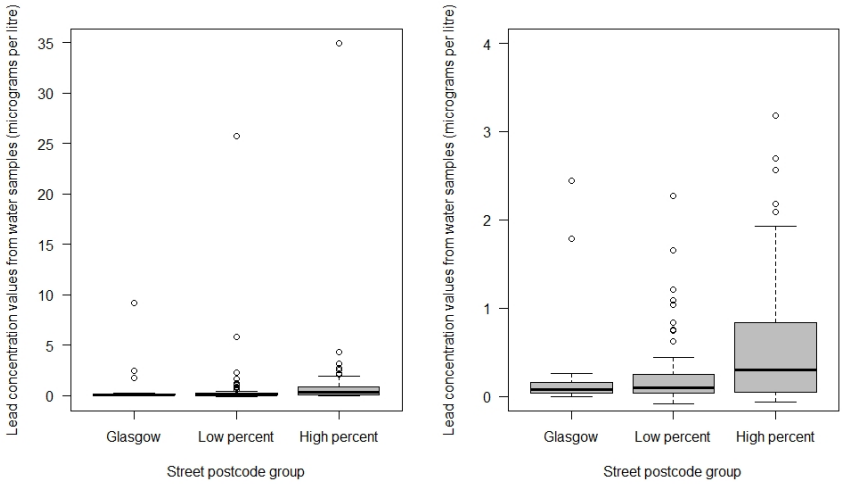

Figure 30: Boxplots showing lead concentration values (in μg/l) from water samples collected from postcodes selected as part of the stratified sampling regime (from postcodes in Glasgow and from postcodes predicted to have high (>=10%) or low (<10%) percentage of houses with lead piping) for a) full range of values and b) values restricted between 0 and 4 μg/l (n = 155)

13 Second round of sampling

The first round of sampling showed that our model was able to predict street postcodes with higher number of houses with lead piping reasonably accurately. However it was less accurate at predicting percentage of houses with lead piping in postcodes which had a low number of houses with lead piping. Model predicted tended to underestimate the percentage of houses in these postcodes with lead piping. The model also underestimated the percentage of houses with lead piping in Glasgow (although this may be due to limited number of samples returned from Glasgow). To further test the accuracy of the model in postcodes with low predicted numbers of houses with lead piping and postcodes in the Glasgow area we undertook a second round of lead sampling.

A further three hundred sampling kits were sent out to postcodes which were selected using the following sampling regime:

- 116 sample kits were sent to postcodes which were marginal in terms of inclusion in the original hurdle model (which was fitted using 308 postcodes). These postcodes had data for all explanatory variables, >=10% of houses had age data, all had >=10 houses according to the latest census data, but only >=6% or >=9% had houses from which lead water samples had been taken by Scottish Water. Taking a further sample from these postcodes would allow the model to be refitted using a larger sample of postcodes. None of these postcodes had been included in the first round of sampling.
- Of the remaining 300 samples, 80 were sent to postcodes randomly selected from the Glasgow area (none of these were included in the previously sampled postcodes and all had >=10 houses available

for sampling). None of these postcodes were included in the original model, as only one postcode in Glasgow was included in the model and this had already been sampled in the first round. Only 6 postcodes selected from Glasgow were estimated to have >=10% of houses with lead piping.

- 52 modelled and 52 non-modelled postcodes were randomly selected from remaining available postcodes across Scotland. These were selected from postcodes which were not located on islands, which had >=10 houses in total, which were not in the Glasgow area and in which <10% of houses were estimated to have lead piping.

As before we checked that the 300 postcodes selected were still in use using the Royal Mail's 'Postcode and Address Finder' (`https://www.royalmail.com/business/find-a-postcode`), the 'click to address' website (`https://craftyclicks.co.uk/address-finder-from-postcode/`) and Zoopla's address finding website (`https://www.zoopla.co.uk/postcode-finder/`). One address was selected for sampling from each postcode using address finding websites.

We used data collected in both rounds of sampling to calculate the overall mean number of houses with lead values >=1 μg/l in sampled postcodes and compare this to our expected value of 30 houses (with a standard deviation 27 houses). We identified some disparity between model predictions and observational data collected in the first round of sampling in postcodes from Glasgow and those in which <10% of houses were predicted to have lead piping. We therefore concentrated sampling on houses in these types of postcodes in the second sampling round to further examine the model fit in these areas. We used samples returned from Glasgow and from postcodes in which <10% of houses were expected to have lead piping in the second sampling round to determine how close our model prediction of number of houses with lead values >=1 μg/l in these areas are to actual values. We also selected postcodes to be included in the second round of sampling which were only marginally excluded from our original model due to having less than the minimum number of houses sampled for lead. We re-ran the original model including new samples from these postcodes and made new predictions based on the new larger model.

14 Model validation from second sampling round

We received tap water samples from the second round of sampling from 171 unique postcodes out of 300 street postcodes to which water sampling kits were sent out. Chemical analysis was carried out on water samples in total (some samples were tested more than once to ensure the consistency and accuracy of the results as the analyses progressed). Where a water sample from a single address within a single postcode was tested more than once we calculated the mean concentration of lead over all samples tested from that address. This gave a total of 171 sample results some of which were means calculated from more than one analysis of a water sample. From this we calculated that approximately 17.5% of postcodes had samples that had a lead concentration of >=1 μg/l (30 out of 171 postcode samples). Of postcodes from which samples were returned, a mean of 7.1% (with standard deviation of 7.7%) positive samples were expected to be returned according to model predictions.

Water samples were returned from 99 postcodes in which <10% of houses were estimated to have lead

piping and 16.2% (n = 16) of these samples had lead concentrations of >=1 μg/l. Of postcodes with an estimated lower percentage of houses with lead pipes for which samples were returned we estimated that a mean of 4.4% (with standard deviation of 1.7%) of houses within each postcode would have lead pipes. Thirty-seven samples were returned from 84 postcodes selected from Glasgow and 10.8% (n = 4) of these had lead concentrations >=1 μg/l compared to a mean of 4.4% of houses within these postcodes (with standard deviation of 3.1%) according to model predictions.

As in the first sampling round, water samples returned from postcodes estimated to have a high percentage of houses with lead piping had higher lead concentration values than postcodes with a low percentage of houses with lead piping, and this difference was found to be almost significant (Mann-Whitney U test: U = 1350.5, p = 0.05, n = 134).

Together with samples collected from the first round of sampling (n = 155; total n = 326) we calculated that 16.6% of postcodes had samples that had a lead concentration of >=1 μg/l (54 out of 326 samples). We expected that a mean of 8.9% (with standard deviation of 9.5%) positive samples would be returned according to model predictions. Approximately 12% of Glasgow postcodes that were included in either sampling round had samples that had a lead concentration of >=1 μg/l (7 out of 59 samples) which is higher than model predictions (mean of 4.7% of houses with lead within these postcodes (with standard deviation of 4.4%)). We expected that a mean of 18.1% (with standard deviation of 11.9%) positive samples would be returned from postcodes with a high percentage of houses with lead piping which was similar to data from both sampling rounds (23.1%). However, we expected that a mean of 4.5% (with standard deviation of 1.8%) positive samples would be returned from postcodes with a low percentage of houses with lead piping which was considerably lower than shown in data from both sampling rounds (14.1%).

Seventy-eight samples were returned from postcodes which were only marginally excluded from the original model (out of 116 postcodes/addresses which were sent samples). We reran the original model including 78 more street postcodes using these samples and new predictions for selected (n = 70,180) and all (n = 76,586) street postcodes based on house age data in each postcode are included in csv files accompanying this report. We validated these new model predictions using data from both rounds of sampling. We found that the 16.6% of sampled postcodes had a lead concentration of >=1 μg/l, while the new model predicted that a mean of 8% (with standard deviation of 9%) positive samples would be returned according to model predictions for these postcodes). Approximately 12% of Glasgow postcodes that were included in either sampling round had samples that had a lead concentration of >=1 μg/l (7 out of 59 samples) which is higher than predictions from the new model (mean of 4.1% of houses with lead within these postcodes (with standard deviation of 4.2%)). We expected that a mean of 17.9% (with standard deviation of 11.6%) positive samples would be returned from postcodes with a high percentage of houses with lead piping which was similar to data from both sampling rounds (23.1%). However, we expected that a mean of 4.3% (with standard deviation of 2.2%) positive samples would be returned from postcodes with a low percentage of houses with lead piping which was considerably lower than shown in data from both sampling rounds (14.1%).

Our model validation using new sampling data confirmed what was found in the first round of sampling.

Analysis of new data from both rounds of sampling showed that model predictions were relatively accurate for postcodes which were estimated to have >=10% of houses with lead piping, but were less accurate for postcodes estimated to have <=10% of houses with lead piping and for postcodes in the Glasgow area. In these postcodes, our model overestimated the number of houses with lead piping. However despite our carrying out two rounds of sampling, we still only have samples from 326 postcodes with which to validate our model. It is possible that further sampling will give results closer to those predicted by the model. We refitted our model using new data collected from sampling and provide new csv files of model predictions for postcodes across Scotland (one where predictions were made for all postcodes with explanatory variable data and another including only postcodes in which house age was available for >=10% of houses). Comparisons with data collected in both sampling rounds showed that the new model also underestimated number of houses with lead piping in Glasgow and in postcodes expected to have low percentages of houses with lead piping.

15 Limitations of model and extensions

There are several caveats and possible improvements that should be considered when applying the model to real data:

- The model was fitted using 308 street postcodes but was applied to over 70,000 postcodes across Scotland. This amount of extrapolation, while necessary given the available data, may lead to inaccuracies in model predictions as the 308 postcodes for which all data were available may not fully represent all postcodes across the country (although efforts were made to ensure that postcodes used to fit the model were as representative as possible). A further 78 postcodes from sampling was included in the model and the model was re-fitted, but due to the relatively small number of new samples this did not significantly change model predictions.

- Spatial autocorrelation was not accounted for in the model due to statistical complexities involved in creating a spatial hurdle model. Applying a spatial component to this type of model is at cutting edge of statistical research and would require further research to implement. However, applying a spatial component to our model may further increase the accuracy of our predictions.

- There were a limited number of data points per postcode, meaning that aggregate lead concentration and house age estimates in each postcode were associated with a high degree of uncertainty. We did not model the effect that this additional uncertainty might have on the results. As above, incorporating this additional uncertainty would require further cutting edge statistical research to implement.

- Only two explanatory variables were included in the model. Other variables not yet identified may be important in predicting number of houses per postcode with internal lead piping (e.g. housing type (Council housing; Housing Association; privately owned houses))

- Data on house age from SAA data sources were missing for approximately half the postcodes in Scotland. House age data from the SHCS could be also included, but this is unlikely to significantly increase the number of postcodes with age data as the SHCS only has data from 15,000 street postcodes

many of which are already represented in the SAA data. The total number of street postcodes in Scotland in 2019 was approximately 144,000.

- Model validation was only carried out using water samples from 326 postcodes. While this gives us an idea of how well the model predictions fit real data overall, more samples would be needed to comprehensively test the accuracy of model predictions in each postcode. Also, only a single address was sampled in each postcode, therefore variation in number of houses with lead piping per postcode was not represented using this sampling method.

Improvements made to the model (such as accounting for autocorrelation and including more data) may increase its accuracy and allow us to apply model predictions to a larger area of Scotland.

Future work:

- Revalidate model using new data

Another way of testing our model would be to use a subset of data available from Scottish Water and other data sources that was not used to fit the model, but which provide an opportunity to test the model on new data. This may be a useful exercise to further validate our model using previously untested data sources.

- Testing source of lead in water samples

Another possible study area is to examine the source of lead found in tap water samples. Chemical analysis of water samples would enable us to identify the geographical source of lead in the sample which may help determine whether lead in tap water has its source in lead piping within a property or in the environment.

References

[1] Zeileis, A., Kleiber, C. Jackman, S. 2008, 'Regression models for count data in R', *Journal of Statistical Software* **27**(8).

16 Appendix A: Literature review summary

The factors affecting lead (Pb) concentrations in tap water in Scotland are summarised and discussed. In spite of the many steps taken to minimise human exposure, Pb is still considered to be one of the most important environmental pollutants, with detrimental health effects especially for foetuses and young children. Although regulatory limits have gradually tightened over the past decades, tap water remains a major pathway of Pb exposure and so the current EU limit of 10 μg L-1 is set to decrease further to 5 μg L-1 in 2029. The main factors which affect Pb concentration in tap water are: (a) presence of Pb-bearing piping and plumbing materials, (b) water chemistry, and (c) other parameters such as water temperature. In Scotland, elevated Pb concentration in tap water is grossly attributable to Pb-bearing piping/plumbing components in combination with the naturally soft and acidic water. For this reason, the natural chemical characteristics of water, and particularly pH and alkalinity, which play a major role in Pb releases into tap water, are controlled to minimise plumbosolvency (i.e. Pb leaching from piping/plumbing materials into water). For example, Scottish Water, Scotland's publicly owned water supplier, elevates the water pH to between 7.3 - 7.8, in accordance with global practice. Furthermore, corrosion inhibitors, such as orthophosphate, are added into drinking water systems to reduce Pb releases to tap water. Orthophosphate addition is widely adopted as a Pb control strategy in Scotland and as a result tap water is, by and large, in compliance with the current limit of 10 μg L-1 (e.g. 99.91% compliance in 2018). To address the root of the problem, i.e. existence of Pb-bearing piping and plumbing materials, Scottish Water has replaced already almost all (98 - 99%) communication pipes. In order to comply with the new drinking water limits, the most promising strategy is the full replacement of the lead service lines (LSL), i.e. both the communication and supply pipes, with the latter being the responsibility of the homeowners. However, the replacement of supply pipes can be costly (up to £3,440) and homeowners can be reluctant to bear this cost. Thus, without financial incentives, these may still remain in place. Importantly, even in cases of full LSL replacement, Pb-bearing materials in house plumbing, such as lead-tin solders, taps, and brass fittings can result in Pb releases to tap water. Therefore, replacement of all Pb-bearing internal pipes and taps, in addition to full LSL replacement, may be required to minimise Pb concentrations in tap water. The age of a building may be a good indicator of the material used for internal piping and plumbing. The use of Pb pipes in house plumbing was banned in 1969 in the UK and so houses constructed before 1970 potentially contain Pb piping. The extent of the problem in Scotland will therefore be related to the age of its building stock. In Scotland a large boom in the new house building activity of the private sector took place after the end of WWII up until the 1970s and so many properties may still be served by Pb-bearing supply pipes. Housing age data is currently not available for the whole of Scotland and is something that would be extremely valuable for future research in order to estimate the number of houses that are likely to contain lead piping or storage tanks.

Full version of Literature Review is available in accompanying document ('Literature review - with summary-final')

17 Appendix B: Tap water analysis summary

The aim of water analysis was to find a postal solution that maximised the return rate for samples from the postcodes selected to test the model. This type of blind study previously used tends only to get around a 30 % return rate as a good outcome. Therefore, it is important that the packages were designed to be easy to understand what was being asked for, why and easy to carry out the request. We considered various barriers e.g. producing photographic as well as written instructions.

The packages were sent out in an A4 Brown envelope containing: Covering letter, 1 x 300 ml Bottle and lid, clear polythene snap lock bag and a plastic postal envelope (with £3 postage and return label attached).

The 2-page covering letter included a written explanation of the study, written instructions and a photographic instruction page to cover an anticipated wide range of recipients understanding. The study partners logos were clearly displayed, and details given of where further information could be obtained. Additionally, a summary of the instructions was printed onto a label and fixed on the bottle. All the envelopes, plastic bags, bottles and bottle lids were numbered (including a discreet number inside the postal envelope) in case of wear or tampering.

The prepared packages were sent out in 2 batches weeks beginning 23rd and 30th January 2020, just ahead of the February schools break (12th-23rd February 2020).

The Scottish Water Lead Analysis Protocol for sample preparation was followed (GIC001 – Analysis of defined elements by Perkin Elmer Nexion 300X ICPMS Spectrometer. WLON-68JM3Y/ed:AL).

Details of analysis are explained in full in accompanying document ('Report˙ICPMS element analysis-200402')