Scotland Water - Lead Piping Data Cleaning

by

Student name:

Student ID:

June 2020

Own Work Declaration

I declare that this my own work, and that no body has collborated, assisted me on this work or any process related to this work.

Contents

Executive summary

Lead is a harmful environment pollution. This project aim at developing a dataset for pinpointing the lead in Scotland water system. It comprises of various datasets including sampling data, water monitoring data, and many more.

The output of this data file is at street postcode level. For each street postcode, we averages data such as lead concentration, pip types, pipe replacement information related to each postcode.

1 Introduction

Lead is one of the most important environmental pollutants and has been shown in several studies to be a trigger for health problems Boskabady et al. (2018). This is because lead poisoning can occur when lead enters the body. In most cases, small amounts of lead can accumulate over time and cause health problems. The health effects of lead poisoning cannot be ignored. Exposure to lead can be harmful, especially to unborn babies and young children who absorb more lead as their bones develop. Signs and symptoms of young children can include irritability and fatigue, loss of appetite and weight loss, abdominal pain, vomiting, and constipation. In addition, exposure to lead from drinking contaminated water can cause illness in adults. High blood pressure, abdominal pain, constipation, joint and muscle pain, numbness or tingling in the limbs, and headache are suspicious symptoms in adults (Water 2020, Scottish Water).

In modern life, the risk of human lead poisoning is very low because lead is now generally not used in paint, petrol or food containers. As a result, most people have very little exposure to lead. However, one of the main potential risks may be drinking tap water. People need to be especially careful if the property has lead pipes, water tanks, or pipes with lead fittings, as this can lead to a polluted water supply. And this problem is taken seriously by Scotland government. The government stipulated that all drinking water should meet the tighter standard for lead (10μg/l) that came into force in December 2013 (Scottish Government). Although most modern water pipes are made of lead-free materials, however, houses built before 1970 may have water pipe which connects the property to the water main in the street that contains lead. Hence, it is the strategic goal of Scottish Water company to replace this part of water supply pipeline and ensure the water quality of customers. And the goal of the company is to increase the reliability, resilience and sustainability of its services to realize lead free water supply. In order to achieve this goal, Scottish Water company is striving to achieve two points for their customers: 1. At least one tap in the property which is supplied with water which has not been in contact with lead pipes. 2. The level of lead in the water is no higher than 3 μg/l.

2 Background

The World Health Organization advise that there is no safe level of lead and the current EU Drinking Water Directive allows a maximum level of 10 μg/l. However, the EU Drinking Water is being revised, with lead levels proposed to reduce to 5 μg/l. And this change is expected in 2019 with 2-year transposition to Scots Law. Hence, the regulatory sampling will be targeted at high risk areas, including the areas of Scottish Water. So, compliance of Scottish Water will drop. However, standards recognize achieving zero lead is unachievable as many brass plumbing fittings are extremely secretive and numerous which contain small amounts of lead. In fact, Scottish Water company has made many efforts and achieved certain results before this. The following is a summary of the information available from Scottish Water company in their treatment of lead pollution in water. Starting with regulations, the company takes steps to meet the standards. Phosphate is added to water to reduce plumbosolvency at 85 WTW. The phosphate reacts with lead pipes to form a chemical layer which reduces the dissolution of lead into the drinking water. The amount of phosphate has been optimized for each zone to obtain lowest lead level at the minimum possible dose based on performance of lead test rigs. The majority of phosphate dosing equipment was installed between 1996 and 2006 in response to the lead standards of 50μg/l to 25μg/l. The measures taken by the company are effective. The 2012 Statistical survey indicated that on average 4% of the 1.8 million communications pipes were lead (72,000) (90% confidence level and + 10% accuracy). The levels varied by zone from 0% to 37%. And the surveys of customer incoming supply pipe indicated a similar percentage of lead but not always aligned. Also, communications pipes can supply multiple properties (2.6 million properties). The random, statistical, regulatory sampling from all property ages indicates that:

1. over 97% of samples had lead levels less than 5 μg/l.

2. 1.5% of samples were between 5 and 10 μg/l (27,000 connections).

3. 1.0% of samples were over 10 μg/l (18,000 connections).

However, it is not enough to meet the latest regulatory standards. There are two reasons. For Scottish Water company, the reason of lead in water in the area of Scottish Water is mainly because of lead service pipes. Actually, Scottish Water treated water is naturally low in lead since they do the water treating in Water Treatment Works (WTW) before tap water transmission. So, the lead is due to contamination by the pipe network and plumbing, the main causes are: 1. The chemical layer can quickly destabilize with changes in available phosphate in the water. 2. Forming the layer on old pipes can take much longer than new pipes (as used in test rigs) due to their condition. 3. Performance may be harder to achieve with soft organic laden waters. 4. The raw materials for phosphate are finite and used by other industries.

For customers, the reason of lead in water in the area of Scottish Water are mainly because of less awareness of lead from customers and the plumbing in their home: 1. Most of SW's customers believe that there are no/very few lead pipes still present in the system. 2. Under 30's has little awareness of lead in the environment – too young to remember move to unleaded petrol. 3. Lead pipes are a 'forgotten subject' little about it in the trade press. 4. The supply pipes, plumbing & storage tanks, solder & joints and brass fittings which contains lead can be used by customers. In conclusion, the places where Scottish Water company need to improve and communicates with customers are: 1. Lead communications (Scottish Water) and or supply pipes (Customer). 2. Lead plumbing & storage tanks (Customer). 3. Lead solder & joints (Both). 4. Brass fittings (Both).

To this end, Scottish Water company has prepared the following plan to further reduce the level of lead pollution in the water. 1. Increasing the knowledge of the levels and location of lead within the water supply system through increased sampling. 2. Informing customers about lead in drinking water – raising awareness & helping them to reduce their risk of exposure. 3. Optimizing phosphate dosing – seeking to achieve the lowest levels of lead in drinking water possible at customers taps, until lead pipe removal allows dosing to be stopped. 4. Remove

lead pipes from the public network – using reactive and planned programs to minimize risk and support the switch off of phosphate dosing. 5. Ensuring customer pipes in contact with drinking water are removed - work with customers and stakeholders to encourage improvements to their supply pipes and plumbing to ensure they have access to lead free drinking water.

3   Exploratory & initial data analysis

3.1   Raw data

When exploring the dataset, the first question that need to be concerned is where does the data come from? The sources of data are mainly divided into three categories according to their uses, which are respectively: 1. Data sets related to the identification of lead concentrations in local water pipes. 2. Data sets related to the identification of phosphates in local water pipes to reduce lead concentrations. 3. Collect data sets of user information, street information and a series of public facilities information within the scope of Scottish Water. The following is a brief description of the source and functionality of each dataset by category.

1. Datasets related to the identification of lead concentrations in local water pipes.

   SW - Lead Comm Pipe Replacements (2004-2018).csv. This dataset contains the information about the completeness status of the communication pipes for each street postcode. For feature 'Wo Create Date' it records the date of replacing the lead Comm Pipe. For feature 'Wo Completed Status' it records the outcome of working status of replacing lead pipes. For feature 'Street postcode' it records the street postcode which the lead Communication pipe belongs to. For feature 'Ads Water Operational Area' it records the address of water operation area. For feature 'Ads Water Supply Zone' it records the address of water supply zone.

   SW - Comm pipe data.xls. This dataset contains the detailed information of the communication pipes of Scottish Water company. For the feature 'Street postcode' it records the street postcode which the lead Communication pipe belongs to. For the feature 'AR10_PROPERTYID' it records the unique representation of each property. For the feature 'Property Type' it records the type of property serviced by communication pipe. For the feature 'Is the property age pre 1970?' it records whether the property age is before 1970. For the feature 'Pipe Material' it records the type of material which the communication pipe is made of as at date of inspection. For the feature 'Date' it records the date of pipe inspection.

2. Datasets related to the identification of phosphates in local water pipes to reduce lead concentrations.

   SW - Scottish Water Zonal Phosphate Levels.xls. This dataset contains the descriptive data on how the releasing of phosphate adapt environments in various WOA regions. For feature 'Sample Date', it records the date of sampling. For feature 'Hydrogen ion', it records the amount of hydrogen ion dosing in drinking water at Water Treatment Zone. For feature 'Lead', it records the amount of lead dosing in drinking water at Water Treatment Zone. For feature 'Phosphorus', it records the amount of orthophosphate dosing in drinking water at Water Treatment Zone. For feature 'Temperature', it records the 'Temperature at Water Treatment Zone'. For feature 'Rig', it records the information of the WOA name.

   SW - Phosphate Dosing WTWs Y or N.xlsx. This dataset describes the information of whether or not orthophosphate dosing of drinking water occurs in water treatment zone. The feature 'WTW Name (Code)' can be used as a connect column of this dataset with the dataset 'SW - All Lead WQ Samples (2010-18).xls'. Its meaning is the name of Water Treatment Zone and it is same as the meaning of Water Treatment Area Name in dataset 'SW - All Lead WQ Samples (2010-18).xls'. The feature 'Phosphate Dosing on Site?' records whether or not orthophosphate dosing of drinking water occurs in water treatment zone.

   SW - All Lead WQ Samples (2010-18).xls. This dataset mainly contains the water quality detection information of the lead poisoning in the area of Scottish Water company. For the feature 'Sample Date Timestamp' it records the sample collection date and time. For the feature 'Result Numeric Entry' it records the lead concentration in tap water (in μg/l).

For the feature 'DMA Name and ID' it records the District Meter Area Name and ID. For the feature 'RSZ Name and ID' it records the Regulatory Supply Zone Name and ID. For the feature 'RSZ Water System Name and Id' it records the Water Treatment Area Name and Id. For the feature 'Sample Date' it records the Sample collection date. For the feature 'Street Postcode' it records the Street Postcode. For the feature 'WOA Name and Id' it records the name and id of water Operational Area. For the feature 'WSZ Name and Id' it records the name and id of water Supply Zone.

3. Collect datasets of user information, street information and a series of public facilities information within the scope of Scottish Water.

Other - UK-HPI-full-file-2019-03.csv. This dataset is about the UK House Price Index (UK HPI) captures changes in the value of residential properties. The UK HPI uses sales data collected on residential housing transactions, whether for cash or with a mortgage. Properties have been included: in England and Wales since January 1995 in Scotland since January 2004 in Northern Ireland since January 2005. Data is available at a national and regional level, as well as counties, local authorities and London boroughs. In this dataset, the useful features are 'Date', 'RegionName', 'CouncilArea2018Code' and 'AveragePrice'. For the 'Date', it records every sample's time. The period of each 'Date' sample is per month. For 'RegionName', it record each name of the selected region. The target region of Scottish Water is contained by some of the region name. And we only need to select them out when merging. For 'CouncilArea2018Code', it is an individual identifier of each 'RegionName'. The usage of it is to merge this dataset by this column with the same column in the dataset 'Other - Postcode_ household count_ urban class.csv'. For 'AveragePrice', it records the average house price in region in each date in monthly period.

Other - SAA_PropertyAgeData.csv. This dataset is about the UK property year recording. The features are 'UPRN', 'Postcode', 'Building_Type', 'Age_Year', 'Age_Category', 'XCO-ORD', 'YCOORD'. For 'UPRN', it stands for Unique Property Reference Number and was created by the Ordnance Survey (OS) from the local UK government. It consists of numbers of up to 12 digits in length and it has a unique number for each land or property. For 'Postcode', it is the street postcode which can be used to merge with other datasets. For 'Building_Type', it records the type of the property. This feature might be a useful predictor with % of the detached buildings in a postcode count potentially. For 'Age_Year' and 'Age_Category', they record the year of house built. It should be noted that these two columns of data are not a complete integer type, and there are other descriptive data that need to be classified manually. For 'XCOORD' and 'YCOORD', they are the coordinates from Digimap projection system.

Other - Postcode_ household count_ urban class.csv. This dataset is about the Scottish Government Urban Rural Classification's information for each street postcode we have in the region of Scottish Water. The useful features in this dataset are 'Street postcode', 'CouncilArea2018Code', 'CensusHouseholdCount2011' and 'CensusPopulationCount2011'. For 'Street postcode' and 'CouncilArea2018Code', the usage of them are to merge this dataset by this column with the same column in the dataset we introduced above. For 'CensusHouseholdCount2011' and 'CensusPopulationCount2011' they record the number of households and population in each street postcode based on Scottish 2011 census.

SW - Postcodes linked to SW Zonal Structure.xlsb. This dataset is an association table that connect the street code with its corresponding WOA, RSZ, WSZ and DMA region. For features with 'WOA', they record the ID and name of each Water Operational Area connecting with postcode. For features with 'RSZ', it records the ID and name of each Regulatory Supply Zone connecting with postcode. For features with 'WSZ', it records the ID and name of each Water Supply Zone connecting with postcode. features with 'DMA', it records the ID and name of each District Meter Area connecting with postcode. Also, the feature 'Total properties' records the number of properties in each DMA and the feature 'Count' records the number of DMA in each postcode.

3.2   Explorotary data analysis

For the dataset used in 'SW - Scottish Water Zonal Phosphate Levels.xls', it is very hard to use. First, it is not a relational dataset, so it is hard to directly merge it with other dataset in the use of Water Operational Area (WOA) name or other identification letter. Second, the dataset is detached, which means the dataset is separate by the different areas with orientation south, east, west and north. Third, the separate part of the dataset has the WOA name recorded hazily, which means the name of each street cannot exactly match the street names in the other datasets. So, in this dataset, first the name of each WOA should extracted by region and the key part of the string subtracted from the name should intercepted for a fuzzy search. To do this, we do not need the region separation so we combine the geographical WOA names of each region into a single column to make it become a relational dataset that can be more user-friendly. Second, we find the WOA name by converting each WOA name to a string and do slicing according to certain rules. The example format of each WOA name is 'XX - Lead rig – WOA name (alias) Zone (name of Pumping Station)'. We should only use the WOA name and slice this part as our merging part. In this procedure, we find that the signs '/', 'A, B, C', '()' are too detailed and does not have these characters in another associated table. So, we ignore those signs in the WOA name merge. We find there are only 99 WOA in the whole region of Scottish Water company. We find the feature 'Total organic carbon' and 'Optimization status change' has too less data and they are meaningless in the research, so we directly delete them and do not obtain them in the merged dataset. Also, the feature 'Sample Comments' are tanglesome. We view this feature one by one and find it is not reasonable, so we drop it from the merged dataset. Another problem we find is the WOA name in 'SW - Postcodes linked to SW Zonal Structure.xlsb' we are merging for is overrated. Some of the WOA name in the 'SW - Scottish Water Zonal Phosphate Levels.xls' dataset has several small subdivisions. In this case, we calculate the mean value of each overrated value matching to the subdivision part. After screening, we also found that some sample points of data were obviously noise points (as shown in the figure below) and some sample points were missing due to too late screening time to enter data (as shown in the figure below). In the case of noise point, we use the method of moving median to detect and delete the noise points. We choose 5 days sample points before each sample points data to calculate the moving median and delete the points are over 20 times than the moving median. In the case of missing value, we directly remove the data points since we need to use the median value of last 5 points, and the missing points are not indispensable.

Note that we use median instead of average value to judge whether a data point collected by rig is an abnormal value, which is because that median is a much more robust value compared to moving average. By experiment, we see that the moving average is greatly affected by the extreme values, and could misclassify some of the normal values as abnormal since the moving average is pulled away by the extreme value from normal values.
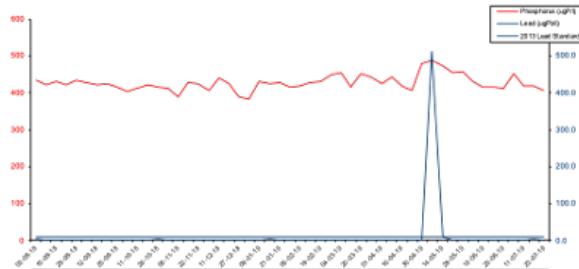

Figure 1: Example 1 of extreme values

Below are examples showing what kind of sample points (in blue) we have deleted from the dataset.

Now the result of doing the above things is one WOA matches one sensor. Then, we need to search the tables for WOA and WSZ at same time by extracting the keywords contains the original WOA name is the dataset 'SW - Scottish Water Zonal Phosphate Levels.xls'. we then
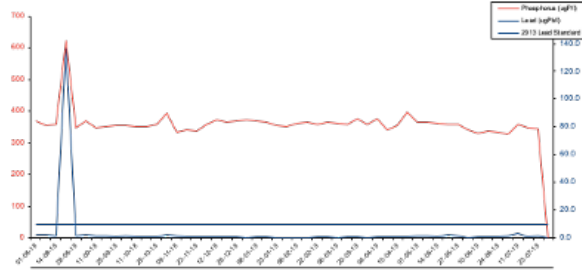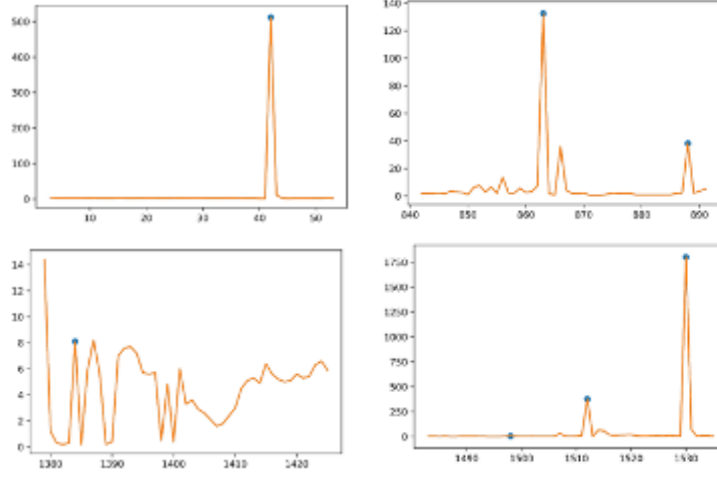
Figure 2: Example 2 of extreme values

Figure 3: Missing values marked by out algorithm

associate the searched keywords with postcode. Hence, such a sensor corresponds to a lot of Postcodes. The existing values collected for each sensor's associated physical coefficients are the average of the last five non-null values. We then do a reverse search to find which sensors correspond to which Postcodes, and calculate the average of the physical parameters of all sensors based on each Postcode. These include: 'Hydrogen ion', 'Lead concentration', 'Phosphorus' and 'Temperature'.

For the dataset used in 'Other - UK-HPI-full-file-2019-03.csv', after we do the initial data analysis work, we make sure that all the average price is for 411 areas region and each of them has the last date price collection at 01/03/2019. Also, each region has at least over one-year collection of samples for data sampling for each month period. Hence, we could make use of the last year period, that is, the 12-months' data sample of each region to find an average region house price of each region in an average year count. After making sure the dataset is as expected as it described, we calculate the average region house price of each area in an average year count. Below is a frequency distribution histogram of the house prices in the region of Scottish Water company. We find that most of the houses are priced between £200,000 and £400,000.

However, we are still not finished. The final goal of this dataset is to calculate the average price base on the distinguishing identifier in the level of the street postcode. Hence, we need to read in another dataset which can be used to connect the two datasets with same feature. After checking at the features dictionary, we find both of the dataset 'Other - Postcode␣ household count␣ urban class.csv' and 'Other - UK-HPI-full-file-2019-03.csv' contains the feature 'CouncilArea2018Code' and the 'Other - Postcode␣ household count␣ urban class.csv' contains the feature 'Street postcode'. Hence, we read in the 'Other - Postcode␣ household count␣ urban class.csv' contains the feature 'Street postcode' dataset and left merge those two datasets with 'CouncilArea2018Code' by left joint. Then we get a table with feature 'Street postcode' numbers multiple mapping with the feature 'AveragePrice' which we select from the original calculated dataset. Hence, we can calculate the average house price for each street postcode. Here is the

Frequency distribution of house price in Scottish Water Region
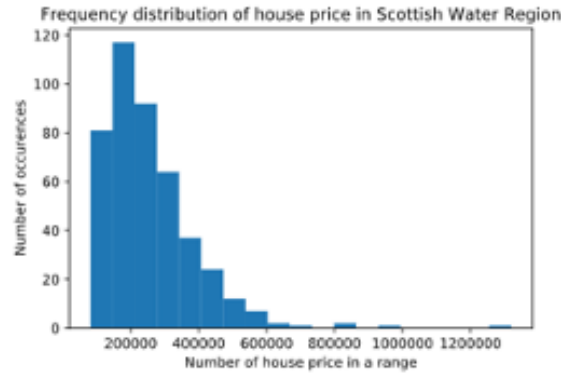
end of this dataset manipulation.

For the dataset used in 'Other - SAA_PropertyAgeData.csv', our goal is to find the percentage of properties built since 1970 according to each postcode. To do this, we first have a view at the dataset and then we find out the features 'Age_Year' and 'Age_Category' are related with our goal. Hence, we find out the datatype of them and find the 'Age_year' is easy to judge since its type is float 64. However, the 'Age_Category' has type object and we find the class of values in the feature 'Age_Category'. After searching and choosing, we find there is a proper class named 'AgeCat: Post 1971' which means the properties with this category value are built after 1970. And this is what we want. By group sort and calculation, finally, we find the ratio of properties built since 1970 according to each postcode.

Frequency distribution of ratio of properties built since 1970
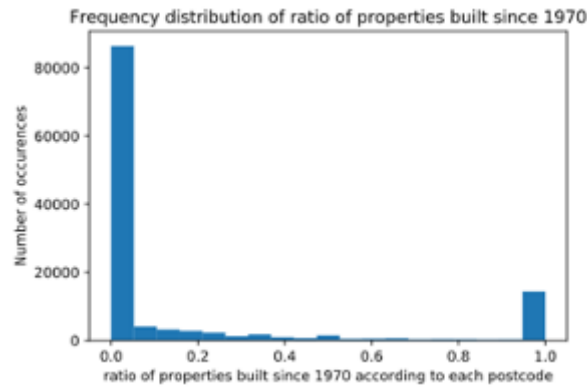
For the dataset used in 'SW - Lead Comm Pipe Replacements (2004-2018).csv', our goal is to find the counts of how many times the pipe of each street postcode has been successfully replaced. Also, within the procedure, we could find whether the street has or has not replace its pipes. To do this, we first read in the dataset an find the information of each feature. Next, we should choose features we need to build a new dataset and make sure all postcodes are capitalized and no blanks in it. Here, we should use the features 'WO Completed Status', 'WO Create Date' and 'Street postcode'. In order to make sure all the replacement of pipes is valid. We should check the replacement date here to ensure the date of replacement is after 1970. To do this, we need to convert the feature 'WO Create Date' to date type, so we can calculate with date type easily. After checking, we find all the samples with dates are after 1970. Then, we check the class types of feature 'WO Completed Status'.


```
All Complete                               5249
All Complete with Comments                 4670
Complete (Partial)                         1434
Cancelled with comments                     664
Bulk Closure                                294
Cancelled by Bulk Closure                    18
System Closure                                2
Name: WO Completed Status, dtype: int64
```

9

Here we find the classes of this feature for the complete status are mainly two types: 'All Complete' and 'All Complete with Comments'. So, we need to generate a column that records the street postcode with whether its pipe has been finished replacement. we name the new feature as 'replacement_finished'. When doing this, we find the elements in 'WO Completed Status' have lots of meaningless blanks and we need to delete them before Boolean slicing. And now we can calculate the ratio of the replacement finished, find whether the street has or has not replace its pipes and count the number of finished replacements for each street postcode. however, the postcode in the above dataset for Scottish Water are not complete, since this dataset only contains the information of street postcode that may has replaced the pipes. We need to use another dataset called 'SW - Postcodes linked to SW Zonal Structure' to find the complete postcode and merge them together. When read in the new dataset we also need to delete the nan value sample points and standardize the postcode format. After let merge, we get a dataset and we need to fill in the NAN values. For the feature 'any_replacement', we replace the NAN value with False. For the feature 'success_replacement_count', we replace the NAN value with 0. In the below table, we could find that most of the street has not replace its pipes.

```
0.0      165767
1.0        5724
2.0        1288
3.0         352
4.0         100
5.0          55
6.0          19
7.0           4
8.0           4
9.0           3
15.0          1
14.0          1
10.0          1
Name: success_replacement_count, dtype: int64
```

References

Boskabady, M., Marefati, N., Farkhondeh, T., Shakeri, F., Farshbaf, A. & Hossein, M. (2018), 'The effect of environmental lead exposure on human health and the contribution of inflammatory mechanisms, a review', *Environment International, IF* **7**(577), 1.

Government, S. (2013), 'Drinking water quality for scotland'.
    URL: *https://dwqr.scot/public-water-supply/drinking-water-quality-faqs/lead/*

Water, S. (2020), 'Lead poisoning'.
    URL: *https://www.nhsinform.scot/illnesses-and-conditions/infections-and-poisoning/lead-poisoning*