Environmental Threads

Carlo Alessi[1], Haoyan Chen[1], Guy Tucker[1], Martin Ivanov[1], Payam Ghoreishi[1]

Abstract
Water is perhaps the most valuable resource on our planet and it is important to monitor its purity. Many water sources on Earth are polluted due to human activity which harm the local flora and fauna. Department for Environment, Food and Rural Affairs (DEFRA) is an government agency that monitors different aspects of water quality and provides datasets for scientist to analysis. Since this data collection has started recently, the best way to use the environmental data is unclear. There are multiple questions which arise: are they monitoring at the right places, is it at the right time, is it frequent enough and most importantly what insights can this process of monitoring can lead to? We answer those questions by doing research on flood and water quality data obtained from monitoring stations located in the Bristol Avon Urban area from 01/07/2016 to 31/12/2016 by conducting data science analysis on the provided data. First the data is narrowed down to information about the Bristol Avon Urban area and several variables are considered including Amonia, Nitrate Nitrogen, and river level. We discovered one strongly positive correlation is found between Oxidised Nitrogen and Nitrate Nitrogen. Therefore, DEFRA is duplicating its data and thus wasting storage and financial resources measuring both variables. The value of one variable can be used to predict the value of the other. This enable the usage of simpler sensors or their reallocation in other places. Furthermore, it is possible to use the observations to predict sensor failure and inaccurate data. Another finding is that flood level at some locations is followed by a peak in Ph or other water qualities, which could be useful for other scientists.

Keywords
Environment — Data Science — Water Quality – River Levels — Correlation

[1] *Department of Computer Science, University of Bristol, Bristol, United Kingdom*

Introduction

In this paper we address the problem of the effectiveness of environmental monitoring. Can environmental monitoring be improved? Can data collection be optimized? What insights can we make from the data. This problem has a large scope so the scope is narrowed down to can the correlations be found in the data collected from stations in the Bristol Avon Urban area. We believe that finding correlations between past measurements can ease the analysis of future investigations and make the monitorings more effective.

1. Related Work

A wide range of works has been carried in the field of environmental data analysis.

Vega et al. [1] analysed the water quality of the Pisuerga river (Duero basin, Spain), to assess how pollution and seasonality affect the quality of the river water, using principal component analysis (PCA) and analysis of variance (ANOVA). They also investigated the individual effects of human activities and climate changes on the river hydrochemistry. Their research focused on a 25 km section of the river, and they used data collected from three stations every three months for two and a half years. The experimental data was normalized to zero mean and unit variance. Then, exploratory data analysis (EDA) was performed by PCA, to assess associations between variables, and hierarchical cluster analysis. Cluster analysis was used to uncover intrinsic structure of the underlying behaviour of the dataset, without making any assumptions about the data, to classify objects or group them into clusters. They used different measures of similarity such as complete linkage, average linkage, and the squared euclidean distance. An ANOVA of the rotated principal components demonstrated that mineral contents are seasonal and climate dependent, and pollution originates mostly from wastewater.

A similar study was conducted by Zhong et al. [2] for the Three Gorges Reservoir Area in China. They presented a big data processing algorithm for water quality analysis, which uses clustering techniques based on fuzzy C-means and hard C-means. The results of the latter were used to improve the performance of the former. In order to asses the effectiveness of their method, they gathered a massive amount of data using the wireless sensor network (WSN) technology with thousands of nodes. They used five classes to discriminate among different levels of water quality, according to the standard of surface water environment quality. Three water parameters were used as features (DO, CODMn, NH_3 - N). Moreover, they assumed that the collected water samples obeyed a uniform distribution. Finally, they reported that their algorithm improved the clustering convergence, and that the features DO and CODMn were highly determinant in

clustering water samples, whereas the NH_3-N parameter was not as discriminating.

Similarly, Aruga et al. [3] investigated surface water pollution phenomena using multivariate statistical techniques. The experimental data was gathered from 31 sampling sites in the Piedmont region, Italy, and consisted of a set of nineteen chemical variables. First, they normalized the data to have zero mean and unit variance. Subsequently, the data was processed using clustering analysis, nonlinear mapping and multivariate feature selection. Moreover, PCA was used to look at the variable correlations and anticorrelations. They concluded that about ten variables out of nineteen were sufficient for the evaluation of the degree of pollution in the rivers.

2. Methodology

We describe and summarize the provided by DEFRA data. Our main objectives include: identify relationships between chemical variables. Then we compare them and discuss the correlations. We identify the difference between variables to evaluate which could be usefull in further research and which not. Then we aim to provide a model forecasting outcomes.

2.1 Software Stack

The chosen software stack is:

- Python (Jupiter Notebook) as a base language,
- Pandas for data manipulation,
- Matplotlib for data visualisation,
- CSVs for data storage.

Python was chosen as the base language due to its high usage in the data science (see Puget [4]) community and it being strongly recommended by lecturers. Pandas was chosen for similar reasons to Python and its excellent community.

For data visualisation Matplotlib was chosen because of its documentation, its support network and seemless intergration with Python. Two other options were considered for data visualisation, Google Charts, and ggplot2. Google Charts is a collection of JavaScript classes developed by Google. It is based on HTML5 and scable vector graphics so no plugins are required and it can create a variety of visualisations including GeoCharts, bubble charts and organisation charts. However it can be time consuming to use the online tools required to produce the charts. ggplot2 is a graphics package for R. Like Matplotlib and Google Charts it too can create a variety of charts including bar plots, violin plots and density plots. However ggplot2 is restricted to developing statistical models.

Data Storage For data storage, several options were considered. These were MySQL, MongoDB, Neo4j and in memory. MySQL is an relational database management system (RDBMS) and is common component in the traditional web application software stack. There are two versions, an enterprise version which contains many additional features including scalability, high availability, and security and an open source version which contains additional features such as clustering and routing. The advantages of using MySQL are the vast amount of documentation available on the website and the support network available in the community. The main disadvantage of using MySQL is the inability to handle spatial data. MongoDB is a document database solution to data storage. Like MySQL, MongoDB comes in enterprise and open source form. The enterprise version contains features such as security and analytics, the open source version is highly scalable and available. The advantages of using MongoDB are it can handle a huge amount of data and requests, it has a highly flexible document storage format, and is capable of complex queries. The disadvantages of using MongoDB is the flexible storage format is susceptible to typos, querying is less flexible and no transaction support. Neo4j is graph database. Like MySQL and MongoDB, Neo4j comes in enterprise and open source formats. The main differences between the two versions is the enterprise version provides advanced monitoring and clustered replication. The advantages of Neo4j are the flexible document storage format, query flexibility and performance and the scalability of the database. The main disadvantage of is Neo4j is it requires learning Cypher, a query language built specifically for Neo4j. Once we evaluated the size of the data and the tasks we want to perform on it, we made the conclusion that a standart relational database MySQL will be sufficient for the purpouse of this project. Furthermore, some of the preprocessing required can be conducted in memory when the size of the data is less than 16 GB (Our modern machines supported memory)

2.2 Data Description and Preprocessing

2.2.1 Data Description

There were two main datasets used in this project, known as **flood** and **water quality** datasets. The flood dataset contains measurements for water level, water flow rate, wind direction and speed, and air temperature. These measurements are provided as a individual values, or as a total, mean, or maximum value of a given period and are taken at different areas including downstream, and at tidal areas and include the date and time taken. This data is provided through a web API with data for the current year available to download in a *daily* CSV, for example on the 05/05/2017 the period of the flood data available will be from 05/05/2016 to 05/05/2017. The daily CSV contains measurements of samples of water. The samples are measured for various chemicals including ammonia and asbestos. Also included the location, date and time the sample was taken and the purpose of the sample. Other samples taken include water temperature, and number of bathers. Like the flood dataset, it is available through a web API or CSV download. CSVs containing yearly data are available for either the whole of England or various areas from 2000 to 2016. CSVs for both datasets also contain API links

to further information about the stations, the measurements and the units of the measurements stored as JSON.

As these datasets contained data for the whole of England, which was determined was too great an area to analyse during this project's timescale. Therefore it was determined to focus on *Bristol and the surrounding area*. The exact area chosen is shown in Figure 1, known as the Bristol Avon Urban operational catchment. This was chosen due to the variety of land types in this area (urban and countryside) and because the authors are all studying at Bristol University. The time period was also aligned so the time period this study focuses on is the second half of 2016 (01/07/2016 to 31/12/2016 inclusive). This is so that the dates of the measurements in both datasets can be aligned.

2.2.2 Data Acquisition and Preprocessing

Although the flood and water quality datasets were available by web API, the datasets were downloaded manually. The main reason for this is the APIs had a limit of the amount of ten thousand items that can be downloaded at once. When daily flood data contains over 200,000 recordings (the flood data for 10/03/2017 contains 237,172 recordings) it would be impractical to write a script to do this when six months flood data was required. Each full CSV of flood data for the second half of 2016 was downloaded from the flood monitoring API archive available from [5]. The water quality monitoring and compliance data for 2016 was downloaded from water quality archive available at [6]. Only the Wessex and Shropshire, Herefordshire, Worcestershire and Gloucestershire water quality datasets were downloaded as these were the smallest datasets that cover the considered area. The total amount of data downloaded amounts to 10.5 GB, 10.4 GB of flood data and 55 MB of water quality data.

The main data preprocessing undertaken was to filter the data so that the only measurements included are measurements from the area given in Figure 1 by following the procedure outlined below.

1. Find the locations of the stations that the measurements were taken from.
2. Map the stations to a water body.
3. Assign each measurement to a water body.
4. Filter the measurements by operational catchment.

Finding the locations of the stations first required finding the number of unique stations. This was done by loading each CSV into memory simultaneously, converting the contents of the CSVs to two pandas DataFrames (df or frame). One frame for flood data and one for water quality Data. The unique stations in these frames were found and put into a separate a frame, only that contained the URLs of the unique stations. The locations were determined by sending requests to the stored using the Python library, urllib, converting the response to json objects and storing the latitude, longitude, northing, easting and national grid reference (NGR) of the station if

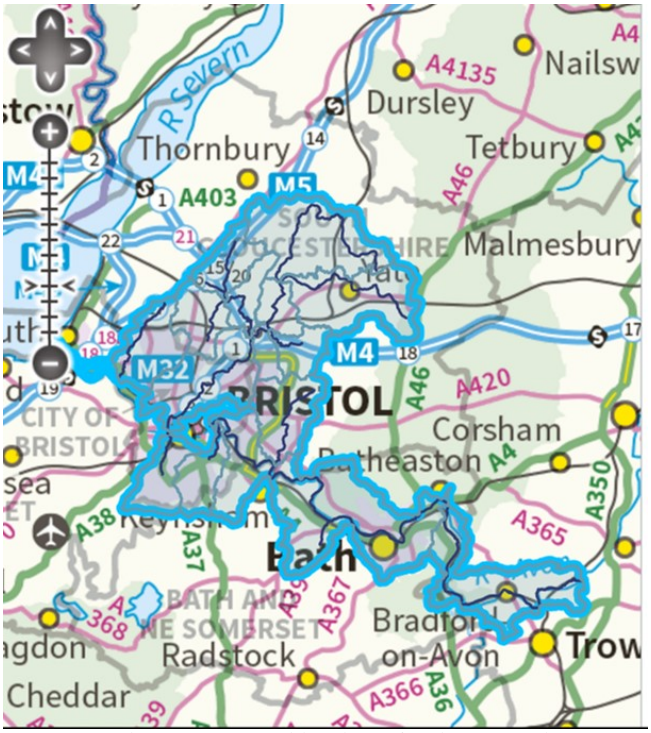


Figure 1. The area covered by the Bristol Avon Urban operational catchment.

they were contained in the response. Out of 5,455 stations only four were missing locations. These are shown in Table 1. The locations of the stations were estimated to ensure no data was removed incorrectly from the area. The estimations were done by typing the label of the station into Google and using the assumptions that a station would be labeled to by the name of it's location and if the station contained RL (the authors believe this stands for river lock) in the label it was close to a river lock.

Mapping the stations to water bodies was difficult. To the authors knowledge there is no web service that converts coordinates to the water bodies they are on and a limited number of stations contained the water body they were on. Therefore extra data and an algorithm was required. The extra data used was the Environment Agency catchment data available from [7]. The classification data stored as CSVs for each of the ten River Basin Districts was downloaded. This data includes the waterbodies, their operational catchment and central coordinates in northings, eastings and NGRs.

The algorithm used is a clustering technique. Using the waterbodies locations as centers, the euclidean distance between the stations and the waterbodies was computed using northings and eastings and the station was assigned to the waterbody it was closest to. Where northings and eastings were not available for the station, NGRs were converted to northings and easting using a custom build script or if NGRs were not available, latitudes and longitudes were converted to northings and eastings using a script written by Hannah Fry [8] which is available as the bng_latlon Python package

through pip. The results can of the algorithm for one waterbody in the area ”Frome (Brist) - Bradley Bk to conf Floating Hbr” can be seen in Figure 2. The plot does seem to bear a resemblance to with the actual map of the actual waterbody shown in Figure 3.

To assign the measurements to waterbodies, the data frame containing the station data was joined to both data frames containing the flood data and water quality data. The join performed was an inner join on the station code, using the pandas merge command. Where necessary the station code was computed in the flood and water quality by using the contents after the last backslash in the URL.

Filtering by operational catchment was done using the pandas filtering facilities. The exact filtering performed was to keep the data that had a operational catchment that started with ” Bristol Avon Urban”. The resultant dfs was written to three csv files, one containing the flood data, one containing the water quality, and one containing the stations and which waterbody they had been clustered to.

There a some limitations with this approach. The main limitation is the vast amount of memory required to perform the operations. Even after loading in all the CSVs which requires at least 10.5 GB of memory additional memory is required to perform the operations. The majority of standalone machines would not be capable to hold all the CSVs in memory. Even with a computer with 16 GB of memory, Python memory errors were occurring frequently. Another is execution time is limited by the amount of time it takes for an external API to process a request. During implementation of this method, exceptions were being thrown when sending requests. It was believed the cause of this was too many requests were sent so a delay was incorporated in the code after fifty requests.

2.3 Chemical variables involved

Although the water quality data reported measurements of 97 chemical variables, many of these were monitored only in particular circumstances (or after particular events), and thus they appeared in few samples. Moreover, only 23 variables were present in more than 15 observations. Table 2 reports the summary statistics of only the most frequent variables, which were present in enough observations. The research focused on determining the pairwise correlation between *Ammonia, Oxidised Nitrogen, Nitrate Nitrogen, Nitrite Nitrogen, Orthophosphate, Water pH, and Water Temperature*. These

Station Code	Latitude	Longitude
E14680	51.689618	-1.678811
E12020	51.544455	-0.830858
E14250	50.727435	-1.598846
2202	52.778884	-3.493577

Table 1. The stations that were missing locations and their estimated latitude and longitude.

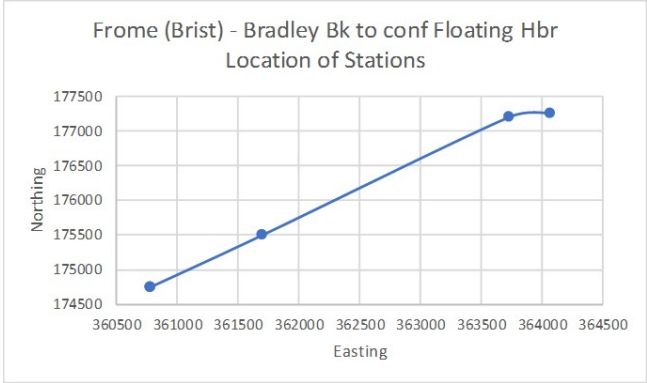


Figure 2. A regression line through the stations found to be in ”Frome (Brist) - Bradley Bk to conf Floating Hbr” waterbody.

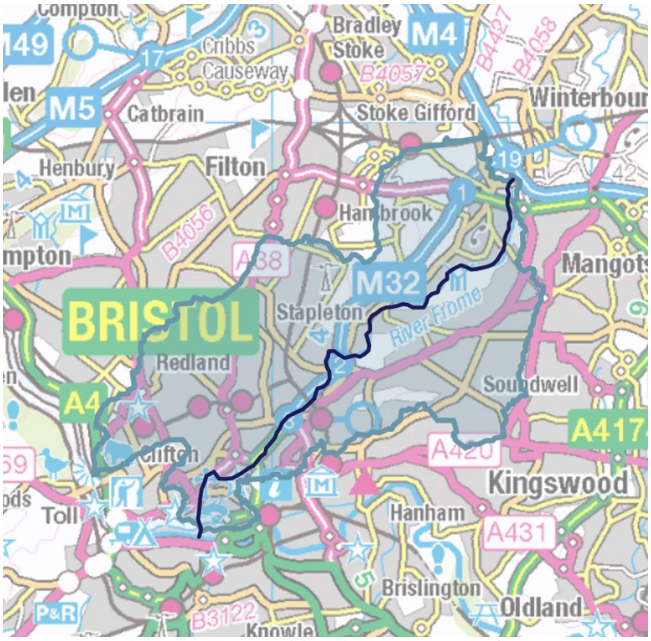


Figure 3. The area covered by the ”Frome (Brist) - Bradley Bk to conf Floating Hbr” waterbody.

chemical variables formed the features of the statistical model.

2.4 Finding a correlation between chemical variables

Correlation is a statistical technique used to measure and describe the strength, direction and shape of the relationship between two variables. Correlation is widely used to make predictions, and requires two scores of the same individual, in this case the same water sample. In order to meaningfully and accurately measure correlation, the chemical variable measurements were filtered, picking just the pairs of values that were sampled at the same time and place.

There are many correlation coefficients. In this research have been mainly used the Pearson product-moment correlation coefficient (R), and Spearman's rank correlation coefficient (ρ). The former measure the strength and direction of a linear relationship. The latter measures how well the two variables can be described by a monotonic function. Pearson's R (shown in Equation 1) always ranges in the interval [-1, 1]. The direction of the relationship is determined by the sign of R. In a positive relationship, both variables tend to move in the same direction (i.e. as one variable increases or decreases, the other one follows the same trend, and vice versa). Whereas if the relationship is negative, the variables tend to move in opposite directions. Moreover, the Pearson's R measures the degree of the linear relationship. The relation is perfect when the absolute value of $R = 1$. On the other hand, when two variables are not linearly correlated at all, $R = 0$. However, R does not perform well when the relationship is curvi-linear, and thus other measures were used as well.

$$R(X,Y) = \frac{Cov(X,Y)}{\sigma X \sigma Y} \tag{1}$$

Spearman's ρ assesses monotonic relationships. Like R its value is between -1 and 1, which occur when one variable is a perfect monotonic function of the other. It is defined as the Pearson's R but it is computed on the ranked variables rgX_i, rgY_i rather than on the raw scores. If the function that bounds two variables is non-monotonic, it can be broken in several intervals in which the function is monotonic. Then ρ would be computed on each interval $X_{start,end}$. The final coefficient would be a linear combination of the individual intervals weighted by the length of the interval itself.

Label	Unit	Min	Max	Mean	std	Frequency
Ammonia	mg/l	0.03	9.17	0.58	1.46	60
Oxidised Nitrogen	mg/l	0.62	7.29	2.72	1.86	55
Nitrite Nitrogen	mg/l	0.0	0.21	0.03	0.04	55
Nitrate Nitrogen	mg/l	0.61	7.19	2.69	1.83	55
Water Temperature	°C	6.70	21.6	14.26	2.89	52
Orthophosphate	mg/l	0.06	0.97	0.23	0.16	52
Water pH	pH	6.85	8.36	7.96	0.27	46

Table 2. Statistical descriptive of the water quality data. Seven most frequent variables selected from a set of 97. For each chemical variable is reported the unit of measure, the number of observations, the minimum, maximum, the arithmetic mean, and the standard deviation.

2.5 Finding a correlation between flood and water quality data

The staff of Environment Agency (EA) wants us to find out some relations between different data sources, which may help them to look for potential issues.

Hence, we integrated the flood data with water quality data using many data preprocessing techniques mentioned in section 2.2 Data Description and Preprocessing. We have to note that the EA gains the monitoring data from water quality stations around 5 times a month, and each time they do not detect all the determinands. In other words, the water quality detection is not continuous in time. In addition, we were not able to acquire the historic flood data generated one year ago because of the limitation of data transfer volume on the Environment Agency Real Time flood-monitoring API.

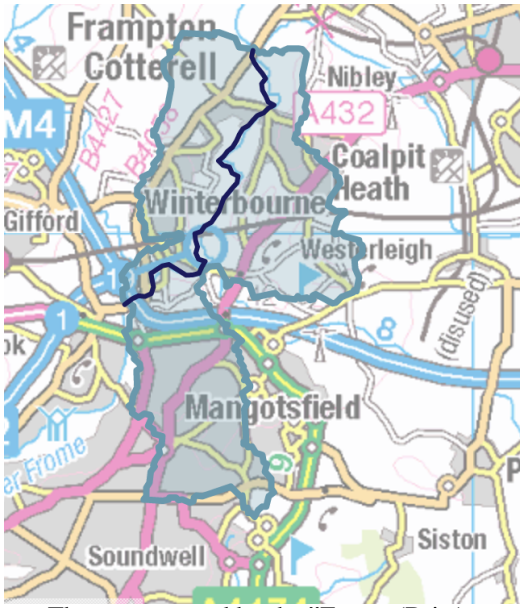
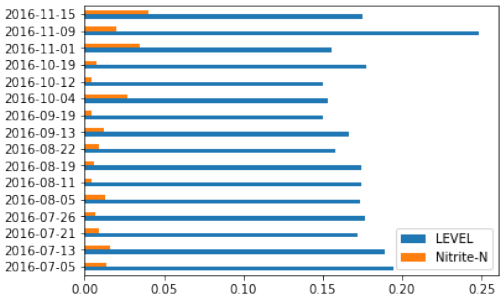
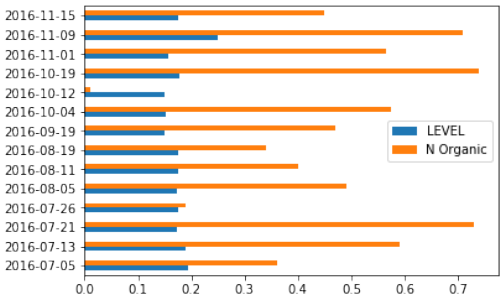


Figure 4. The area covered by the "Frome (Brist) - conf Laddon Bk to conf Folly Bk" water body



(a)



(b)

Figure 5. (5a) Time series data for river level and Nitrite-N. (5b) Time series data for river level and N Organic

As shown in Figure 4, we select an actual area covered by the "Frome (Brist) - conf Laddon Bk to conf Folly Bk" water body (waterbody id: GB109053027820) as a sample area, and we chose two determinands: Nitrite-N and N Organic, which have higher frequency of detection than others. We joined flood data and water quality data by the same waterbody id generated in data preprocessing stage. The flood data set includes river flow data and river level. We analyzed the relation between **river level** and those **determinands** mentioned above. The data we analyzed was from July to December 2016, and two bar charts are shown in Figure 5, which can demonstrate this time series data.

As shown in Figure 6 shows data from particular sensor in bristol area. It suggests that after a peak in water flow level, after a short period there could be a peak in other chemicals. This hypothesis turned out to be true in only few of the locations which we cannot explain further due to lack of data about the geological locations.

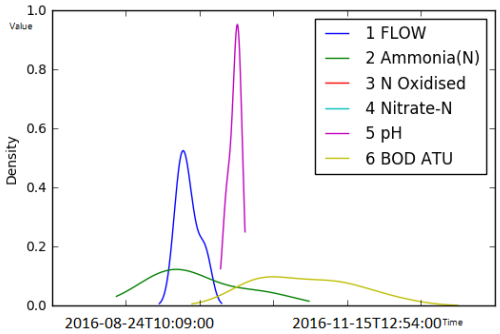


Figure 6. Values observed by a single sensor

Furthermore we tried scientifically find correlation between flood level and chemicals. Spearman's rank correlation coefficient was used to analyze the correlation between river level and two features (Nitrite-N and N Organic) in time series data. As shown in Table 3, there is a 'weak' relation between river level and Nitrite-N, since the coefficient is around 0.21 between 0.20 and 0.39. Although the coefficient between river level and N Organic (round 0.29) is slightly bigger than Nitrite-N's, it is really ambiguous which could be ignored. It was found that there are no relations between flood data and water quality data.

	LEVEL	**Nitrite-N**	**N Organic**
LEVEL	1.000000	0.208791	0.285294
N-Organic	0.208791	1.000000	0.450549
Nitrite-N	0.285294	0.450549	1.000000

Table 3. Spearman's rank correlation analysis

3. Results and Discussion

3.1 Analysis of Linear and Monotonic correlations.

In this section are described the linear and monotonic relationships between the variables. Table 4 shows the Pearson pairwise correlation matrix between the chosen variables. Table 5 reports the Spearman pairwise rank-correlation matrix. Both matrices are symmetric with respect to the descending diagonal because correlation is a symmetric relation ($A\rho B \rightarrow B\rho A$). Moreover, since correlation is a reflexive relation ($A\rho A$), each variable has a perfect relation with itself. Therefore the values that lie on the descending diagonals are all 1s.

As Table 4 shows, there is a strong positive linear relationship between Ammonia and Orthophosphate ($R = 0.73$), although not perfect. The relation between Ammonia and Nitrite Nitrogen is also good ($R = 0.58$). Moreover, Oxidised Nitrogen has a perfect linear relationship with Nitrate Nitrogen ($R = 1$), and a fairly good relation with Nitrite Nitrogen ($R = 0.61$). Conversely Water Temperature was linearly correlated with none of the variables.

The Pearson coefficient has many limitations. It is not appropriate to analyse non-linear relationships. In addition,

	Ammonia	Oxidised Nitrogen	Nitrate Nitrogen	Nitrite Nitrogen	Orthophosphate	Water pH	Water Temperature
Ammonia	1	0.2	0.19	0.58	0.73	-0.47	0.01
Oxidised Nitrogen	-	1	1	0.61	0.35	-0.11	0.08
Nitrate Nitrogen	-	-	1	0.6	0.35	-0.1	0.08
Nitrite Nitrogen	-	-	-	1	0.43	1	-0.04
Orthophosphate	-	-	-	-	1	-0.04	0.03
Water pH	-	-	-	-	-	1	0
Water Temperature	-	-	-	-	-	-	1

Table 4. Pearson's pairwise correlation matrix. The correlation between Oxidised Nitrogen and Nitrate Nitrogen is perfect. Ammonia is strongly correlated with Orthophosphate, but its relationship with Oxidised Nitrogen is weak. Water Temperature is correlated with none of the variables.

z

	Ammonia	Oxidised Nitrogen	Nitrate Nitrogen	Nitrite Nitrogen	Orthophosphate	Water pH	Water Temperature
Ammonia	1	0.37	0.37	0.71	0.65	-0.50	0.21
Oxidised Nitrogen	-	1	1	0.62	0.53	-0.07	-0.02
Nitrate Nitrogen	-	-	1	0.62	0.53	-0.07	-0.03
Nitrite Nitrogen	-	-	-	1	0.58	-0.29	0.06
Orthophosphate	-	-	-	-	1	-0.22	0.22
Water pH	-	-	-	-	-	1	0.04
Water Temperature	-	-	-	-	-	-	1

Table 5. Spearman's pairwise correlation matrix. Ammonia has a strong monotonic relationship with Nitrite Nitrogen and Orthophosphate.

it cannot determine cause-and-effect relationships, and does not distinguish between dependent and independent variables. Moreover, it is negatively affected by the presence of outliers, which impact on the R value and the line of best fit.

The monotonic relation was analysed as well. Spearman rank-order correlation coefficient is much less sensitive to outliers. From the fourth and fifth column of Table 5 it is clear to see that Nitrite Nitrogen and Orthophosphate have a good monotonic relationship ($\rho = 0.58$). Moreover both of them have fine ρ values with Ammonia, Oxidised Nitrogen, Nitrite and Nitrate Nitrogen. The reported values were not enough to assess whether the variables were correlated due to the limitations stated above. The shortcomings of the coefficients were addressed by visualizing the data distributions.

3.2 Finding outliers

The presence of outliers weakened the previous methods. The limitations were addressed visualizing the data distributions through box plots. An outlier is a data point that does not fit the general trend of the data. It can be manipulated or removed as long as the reasons are justified. In this research the outliers were identified and then removed in order to find the line/curve that best fitted the data. There are many ways to find outliers. Our approach considered the Interquartile Range (IQR). The IQR is the difference between the third and first quartile of the data. A data point was defined as outlier when its score was either above the third quartile by $3IQR$ or below the first quartile by $3IQR$. The box plots in Figure 7 visualize the results. The research uncovered that several Ammonia measurements were classified as outliers. Whereas the other variables had either few or no outliers. Removing outliers before computing the correlation coefficients was key to better determine the strengths of the relationships, as well as to delineate the function that might bound two variables.

3.3 Polynomial regression

The data distributions and the trend of the variables were analyzed further. The coefficient index describes the direction and the degree of the relation between two variables. Moreover, it is possible to observe the shape of the relation (if any) through a scatter plot. In a scatter plot, the selected variables populate the x and y-axis to form points in the graph. The relation is linear if the points approximate a straight line, it is curvilinear if they approximate a curved line.

It was assumed that there was a relationship between an independent variable x and a dependent variable y, and that the latter was modeled as an n^{th} degree polynomial in x. Thus polynomial regression was used to fit a nonlinear model to the data. However finding the polynomial that best fitted the data was not trivial. The method used was the Least Squared Error. A limitation of this approach was that it was always possible to achieve zero error. This was because k data points could be perfectly matched by a polynomial of degree $k-1$. A model thus constructed could overfit the examples to the extent that it would not generalize well to unseen data. There were many ways to avoid overfitting. Although cross-validation could be used to find the model that best performed on average, here was proposed a simple yet effective algorithm. The algorithm iterates through the set of possible degrees k, computes the Squared Error (SE), and converges when the following condition is met:

$$|SE(k) - SE(k+1)| < \varepsilon \tag{2}$$

Therefore the polynomial degree was increased until the difference of the Squared Error from one iteration to the next was big enough. The results of the polynomial regression are shown in Figure 8. It was found that when the fitted polynomial was above the first degree the Squared Error was high. Whereas polynomials of the first degree fitted their data with an error non greater than one. However, good polynomial fit not always led to high correlation values. Although the relationship between N Oxidised and Nitrate-N was perfect (8a), the correlation coefficients between Ammonia and pH were small (reffig:s8). The latter was due to the lack of homoscedasticity of the data points (i.e when the variances along the line of best fit remain constant through the line). Therefore it achieved a poor Pearson's R, even though the relation looked linear.

Equation 3 defines Nitrate Nitrogen as a linear function of Nitrate Oxidised. It is really close to the identity function $y = x$ and thus it has two important implications. From a computational perspective, Defra is duplicating its data and thus wasting storage resources. More importantly, Defra is wasting money measuring both variables. The value of one variable can be used to predict the value of the other. This enable the usage of simpler sensors or their reallocation in

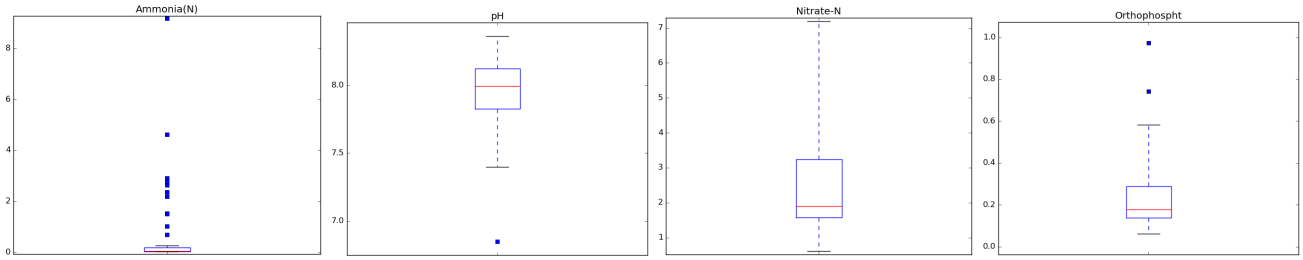


Figure 7. Box-plot of the variables. The top and bottom of the box represent the third and first quartile respectively. The red horizontal line is the median. The top and bottom whiskers span to 3IQR above the third and below first quartile respectively. Ammonia measurements presented many outliers. The pH had one outlier below the first quartile. Orthophospate had two outliers above the third quartile. There were no outilers in the Nitrate measurements.

other places.

$$y = 0.987x + 0.004 \tag{3}$$

Other equations could be defined from the polynomial regression analysis. However the generalization abilities of our models were not tested and takes part of future work.

4. Conclusions

This paper gave insights on the water quality and flood data. Pearson's R and Spearman's ρ were used to determine the degree and direction of linear and monotonic correlation between the following chemical variables: *Ammonia, Oxidised Nitrogen, Nitrate Nitrogen, Nitrite Nitrogen, Orthophosphate, Water pH, and Water Temperature*. Moreover the shape of the relation was assessed observing scatter plots. The presence of outliers in the data harmed the correlation coefficient values. IQR was used to identify and remove the outliers. As a result it was possible to meaningfully fit the data with a polynomial, although often it was not enough to determine the trend of the variables. It was found that the relationship between Oxidised and Nitrate is perfectly linear and monotonic, and is described by Equation 3. Conversely it was not found any strong evidence that other variables are bound by a specific function. This tells us that some of the qualities of water are correlated while others not. This correlation can enable the usage of simpler sensors or relocating existing ones to less monitored locations. The research on time series of water flood showed no direct correlation with the water quality. However, in some of the plots we can observer that this correlation is present but with a distance in time. Thus, further research could confirm that increased level of water precedes a peak in Ph or other qualities of the water. This could be used by other scientists researching water sources and interested in water measurements.

Acknowledgments

The authors would like to thank the Environment Agency for setting the problem, providing the data and their support throughout the project.

References

[1] Marisol Vega, Rafael Pardo, Enrique Barrado, and Luis Debán. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water research*, 32(12):3581–3592, 1998.

[2] Yuanchang Zhong, Liang Zhang, Shaojing Xing, Fachuan Li, and Beili Wan. The big data processing algorithm for water environment monitoring of the three gorges reservoir area. In *Abstract and Applied Analysis*, volume 2014. Hindawi Publishing Corporation, 2014.

[3] Roberto Aruga, Giovanni Negro, and Giorgio Ostacoli. Multivariate data analysis applied to the investigation of river pollution. *Fresenius' journal of analytical chemistry*, 346(10):968–975, 1993.

[4] Jean-Francois Puget. The most popular language for machine learning and data science is . . . , 2017.

[5] Environment Agency. Real time flood monitoring data: Archives, 2017.

[6] Environment Agency. Download open water quality archive datasets, 2017.

[7] Environment Agency. Catchment data search, 2017.

[8] Hannah Fry. Converting latitude and longitude to british national grid, 2012.

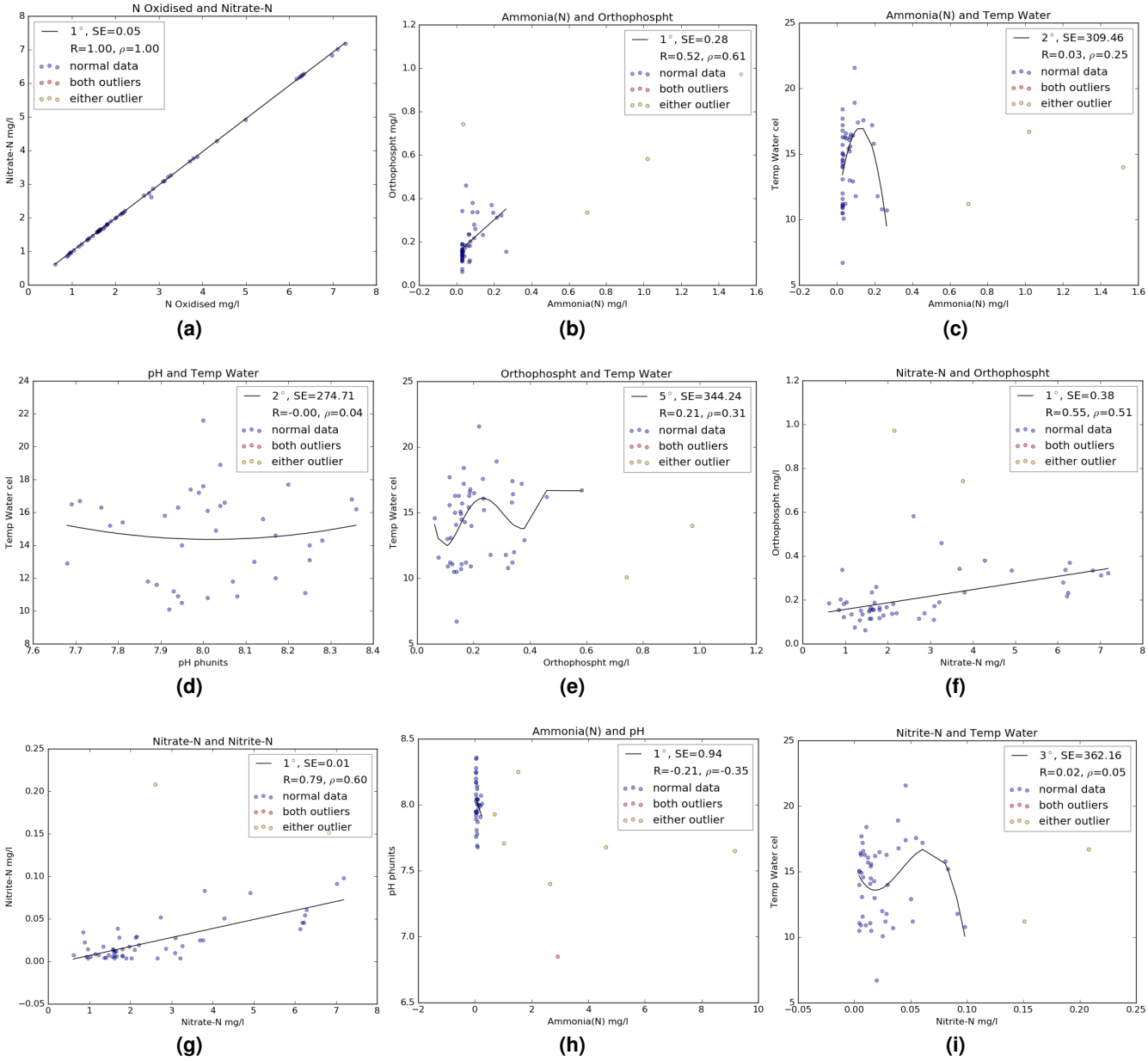


Figure 8. Scatter plots. For each plot is reported the line/curve of best fit. In blue the data cleaned from outliers. In yellow the points where just one variable score was outlier. In red the points where both scores were outliers. (8a)The linear relation between the variables is perfect. (8b) Presence of outliers and good relationship. (8c) Concave polynomial. Both linear and monotonic relationships are poor. (8d) No outliers. No relationship. (8e) Highest degree of polynomial found. High squared error and weak relationship. (8f) Fine linear correlation and small error. (8g) Strong linear correlation and small error. (8h) Poor correlation coefficients and and small error, but non-homoscedasticity of the data led to poor correlation values. (8i) Weak correlation and high error.