Water Quality Monitoring: Analysing data for signs of over-monitoring

GitHub: https://github.com/nadia1123/Applied-Data-Science-Project

May 7, 2017

Abstract

The Environmental Agency collects millions of water quality measurements from thousands of sampling stations around England every year. Maintaining these sampling stations costs money and resources. In this report, we analyse water quality data for the Wessex region over a 5-year period and use correlation measures to identify locations and pollutants that the EA is over-monitoring.

1 Introduction

The Environmental Agency (EA), acting under the Department for Environment, Food & Rural Affairs (DEFRA), collects millions of measurements around England every year. Under the "DEFRA Open Data Strategy" of 2013, the EA now hosts over 2000 datasets covering various aspects of the environment but the focus of this project is on water quality.

In 2016 alone, over 14 million water quality measurements were made of over 200 different determinands. A determinand refers to a specific property of the water which can be measured. For example, it can be a measure of the amount of some chemical contaminant in the water or simply a measurement of water temperature. These water quality measurements are taken for two reasons. The first is the need to look after the environment and ensure that environmental protection is applied where necessary. Secondly, under *Environmental Permitting (England and Wales Regulations) 2010*, companies which discharge liquid waste into the environment are required to have environmental permits for doing so. These compliance measurements are taken to ensure that the permits are not breached.

This project aims to investigate whether water quality measurements correlate with one another, and if so, how. Identifying such correlations would allow us to provide suggestions as to where over-monitoring may be occurring.

If there is some relationship, for instance, between the nitrogen content and the iron content at a sampling station, one could argue that only one of the two measurements is required. Such results could allow the EA to spend less money on monitoring. In this project we wish to identify both pollutants and locations that are over-represented in water quality measurements, investigating whether measurement values can be inferred from other similar measurements.

We do not expect to see large-scale spatial correlations, but it is reasonable to hypothesize that pollutant levels may be correlated along a stretch of river or in the same water basin.

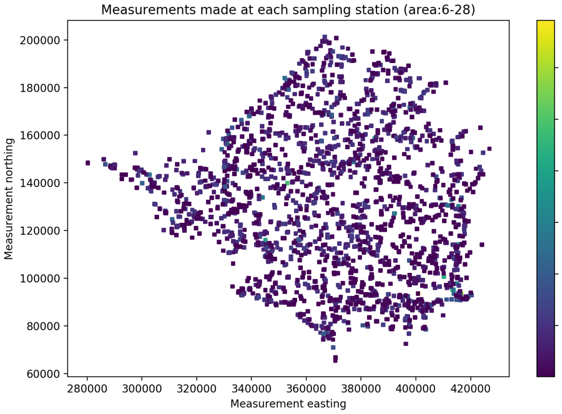


Figure 1: Map of sampling stations in the Wessex area which made measurements in 2001-2016. The colour of each point indicates the number of measurements taken by the corresponding sampling station.

Figure 1 shows the measurements that have been made in the Wessex region from 2001-2016. This region is situated on three separate water basins, therefore we may expect to see different results across the region.

In this project we aim to investigate three main types of correlation to assess whether the EA is over-monitoring water quality:

1. **Temporal correlations** Seasons or other time resolutions (such as months, years, etc.) may exhibit periodicity for the concentrations of specific determinands around the UK, indicating natural changes in water quality over time. Patterns found here could be used to determine whether specific determinands are being measured with appropriate frequency.

2. **Pollutant correlations** We investigate the relationships between determinands at sampling stations where multiple determinands are being monitored. Ideally, if strong correlations are found then we could argue that only one determinands needs to be measured at that sampling station. However due to varying soil contents and rainfall estimates this relationship will be at a local level, rather than global - a 'one size fits all' rule for

pollutant correlation is not feasible when only a single dataset is being considered.[1]

3. **Spatial correlations** We also investigate how measurements correlate in the spatial domain. From this we can determine sampling stations which are not required.

To investigate the correlations between pollutants and between sampling stations, we will obtain sets of pairs of time series consisting of relevant sample measurements and perform correlation analysis. In particular, we perform cross correlation analysis.

2 Initial steps

2.1 The EA Water Quality Archive

The EA provides historical water quality sample measurements from 2000 to 2017 through its Water Quality Archive.

> "Samples are taken from sampling points round the country, including: agricultural, coastal, estuary, rivers, lakes, ponds, canals, sewage discharges, trade discharges, pollution investigation points and waste sites." [1]

The EA provides users with two ways of accessing the data. The data can be downloaded as CSV files. Here, users can choose to download sample measurements from individual regions (eg Wessex, Yorkshire, etc.) or the whole of England for each year. Alternatively, users can make use of the Water Quality API, which provides users with more options.

We accessed the data through both means and each had its own advantages. Downloading the CSV files was useful as it meant that a SQL database could be kept on a local machine. This database could then be queried and return CSV files of interest which could then be used in the iPython notebook. We worked with this database for much of the initial data analysis. However, the API was also very useful, making it much easier to obtain very specific results quickly and not requiring any data storage solutions on our part.

The data contains information on sampling stations and their sample measurements. Sampling station information includes ID and easting and northing coordinates. Sample measurement information includes determinand ID, determinand label, sample date, and result. As the data for the whole of England was too large for our initial data analysis, we focussed on data from the Wessex area. This geographical region includes the county of Bristol. All of the following findings in this report are based on this specific area.

2.2 SQL Database

By using a database we can make fast and flexible queries on a large amount of data. We chose to use the SQL language because the data is tabular. Furthermore, the SQL language meets our complex query needs, allowing us to conveniently merge and split datasets as required. There are many implementations of SQL available; we chose to use **mySQL**.

After downloading the Wessex CSV datasets for the past five years and storing them in our SQL database, we noticed that some data wrangling would be necessary:

- The data includes many attributes, some of which are irrelevant for our needs.

- There is one CSV file for each year, so records for 2012-2016 are spread amongst five different CSV files.

- Some stations do not provide measurements (are not 'active') every year.

- Similarly, not all of the determinands are measured every year.

The final two points raise the issue of data quality, which will be discussed further on. For now we continue our discussion of our use of a SQL database.

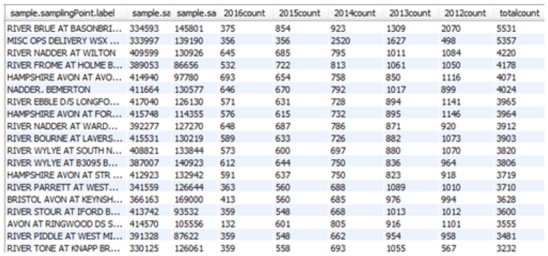
Using SQL, we found the sampling stations which were active in each of the past 5 years (2012-2016), as we want to obtain time series data which does not have any gaps of over a year. This reduces the number of sampling stations that we consider to 978, from an average of roughly 1380 sampling stations per year. We do the same for the determinands.

We also identify the important attributes and ignore the attributes which are irrelevant. Only a few attributes are needed to identify every record; such attributes are called the primary keys. We keep only these primary keys and the attributes which we will use for our further work. For instance, the individual @id attributes of each sample measurement are not useful to us. For our time series data we only need the determinand label and code, result, date, and sampling station, so we can filter the database for these attributes.

We use SQL to obtain basic statistics on the data. We calculated the number of measurements for each station and for each determinand for every year. We then ordered the stations and determinands by their total counts. These measurements help us identify which stations or determinands have more records. **Figure 2** shows a table of sampling stations sorted by total number of sample measurements. In addition to creating a table of sample counts grouped by sampling station, we also create a table of sample counts grouped by determinand, which we use in **Section 2.4.**

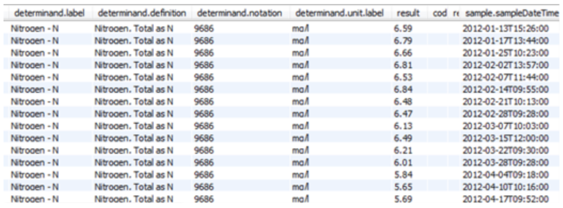
When we perform auto correlation and cross correlation, we will want to obtain measurements from a particular station. To do this, we identify which attributes we want to use, split them from the original table, and finally, merge the records from different years. SQL provides flexible operations to meet these requirements. **Figure 3** presents a table of nitrogen sample measurements.

[1] For this to be feasible other monitored data such as soil quality would need to be considered, but at this level of dimensionality we run the risk of overfitting data.



sample.samplingPoint.label	sample.sa	sample.sa	2016count	2015count	2014count	2013count	2012count	totalcount
RIVER BRUE AT BASONBRI...	334593	145801	375	854	923	1309	2070	5531
MISC OPS DELIVERY WSX ...	333997	139190	356	356	2520	1627	498	5357
RIVER NADDER AT WILTON	409599	130926	645	685	795	1011	1084	4220
RIVER FROME AT HOLME B...	389053	86656	532	722	813	1061	1050	4178
HAMPSHIRE AVON AT AVO...	414940	97780	693	654	758	850	1116	4071
NADDER. BEMERTON	411664	130577	646	670	792	1017	899	4024
RIVER EBBLE D/S LONGFO...	417040	126130	571	631	728	894	1141	3965
HAMPSHIRE AVON AT FOR...	415748	114355	576	615	732	895	1146	3964
RIVER NADDER AT WARD...	392277	127270	648	687	786	871	920	3912
RIVER BOURNE AT LAVERS...	415531	130219	589	633	726	882	1073	3903
RIVER WYLYE AT SOUTH N...	408821	133844	573	600	697	880	1070	3820
RIVER WYLYE AT B3095 B...	387007	140923	612	644	750	836	964	3806
HAMPSHIRE AVON AT STR ...	412923	132942	591	637	750	823	918	3719
RIVER PARRETT AT WEST...	341559	126644	363	560	688	1089	1010	3710
BRISTOL AVON AT KEYNSH...	366163	169000	413	560	685	976	994	3628
RIVER STOUR AT IFORD B...	413742	93532	359	548	668	1013	1012	3600
AVON AT RINGWOOD DS S...	414570	105556	132	601	805	916	1101	3555
RIVER PIDDLE AT WEST MI...	391328	87622	359	548	662	954	958	3481
RIVER TONE AT KNAPP BR...	330125	126061	359	558	693	1055	567	3232

Figure 2: Sorted stations with basic statistics



determinand.label	determinand.definition	determinand.notation	determinand.unit.label	result	cod	r	sample.sampleDateTime
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.59			2012-01-13T15:26:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.79			2012-01-17T13:44:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.66			2012-01-25T10:23:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.81			2012-02-02T13:57:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.53			2012-02-07T11:44:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.84			2012-02-14T09:55:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.48			2012-02-21T10:13:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.47			2012-02-28T09:28:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.13			2012-03-07T10:03:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.49			2012-03-15T12:00:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.21			2012-03-22T09:30:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	6.01			2012-03-28T09:28:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	5.84			2012-04-04T09:18:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	5.65			2012-04-10T10:16:00
Nitrogen - N	Nitrogen. Total as N	9686	mg/l	5.69			2012-04-17T09:52:00

Figure 3: Nitrogen samples from a particular station over 5 years

2.3 API

We also use the API as it allows all group members to obtain specific results quickly. This is useful particularly for the spatial data analysis, as using the API allows us to quickly specify specific latitude and longitude coordinates rather than northing and easting coordinates. The API also has a useful feature which allows us to select all the sampling stations within a given distance of a specified latitude and longitude coordinates.

Using gmaps, the Google Maps plugin for the iPython notebook [2], we can easily create maps visualising the locations of sampling stations. **Figure 4** is a heatmap of sampling stations in Bristol.

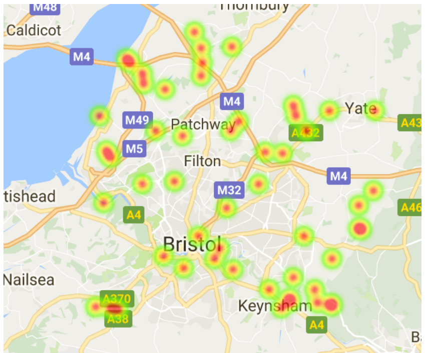


Figure 4: Heatmap of 2016 sampling station locations in Bristol

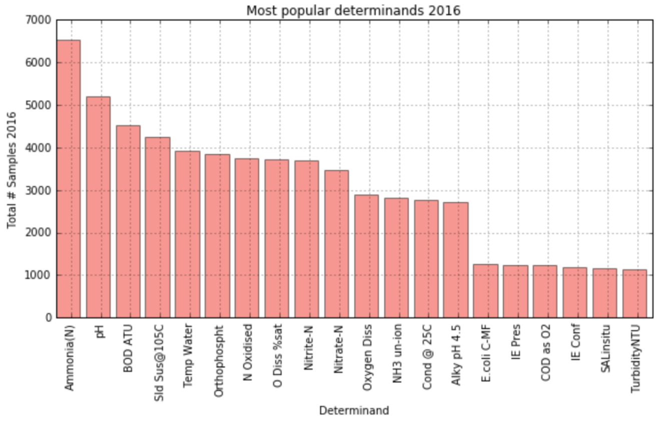


Figure 5: Most popular determinands 2016

2.4 Initial Data Analysis and Data Quality

Initial Data Analysis

We perform initial data analysis by creating custom queries in SQL, downloading the resulting CSV files, and then using them together with the **pandas** module in an iPython notebook. All of the results we discuss below are specific to the Wessex area. We restrict our work to the local area of Wessex for the purposes of this project for convenience, although scaling up would be possible.

An obvious question to ask is, what does the EA measure when it measures water quality? **Figure 5** thus presents the 20 most popular determinands in 2016 and the number of samples collected for each. In total, there were 291 determinands and 89,709 samples made in 2016. Of these 89,709 samples, 61,307 of them were samples of the top 20 determinands. Therefore 68% of all samples in 2016 were samples of these top 20 determinands.

It is also interesting to consider the number of sample measurements made each year. This number appears to have decreased every year since 2012 (159,933 samples) to 2016 (89,709 samples). In total, 655,572 samples were made over the 5 year period.

Data quality

During our initial data exploration, we noticed that some stations listed by the EA are not 'active', and some determinands are not measured at the same frequency between different years. We show this by calculating the frequency of sample measurements. **Figure 6** presents the number of sampling stations in 2016 and the number of days in the year for which they had samples. It is found that over 77% of the sampling stations made 10 or less measurements in 2016.

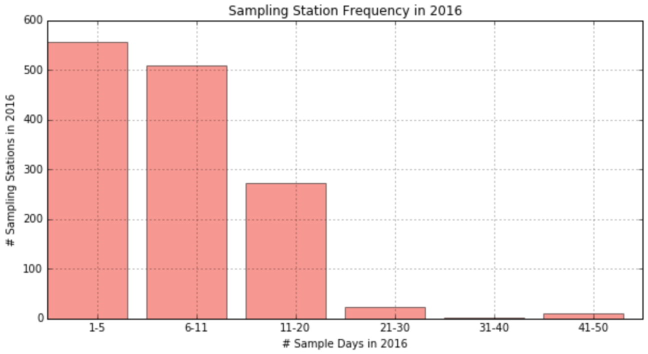


Figure 6: Sampling station frequency in 2016

In later discussions with the EA, this discovery of data sparsity was found to be an area of interest. The EA publishes data as part of DEFRA's "Open Data Strategy" but currently has no strategy for assessing the quality of the data it publishes. It was surprising to see that the majority of sampling stations had sample measurements published for such a small number of days in the year. Furthermore, the API lists many sampling stations as being of 'open' status although they have not provided any sample measurements in the past year. We find the sparsity of the data in the temporal domain to be a major limiting factor in our work.

3 Correlation measures

In order to provide a description of the models we have built in order to achieve our goal, we first offer a brief summary of the correlation measures used.

3.1 Linear correlation

The familiar **Pearson correlation coefficient (PCC)** measures the linear correlation of two time series. The PCC takes values in the range $[-1, 1]$ where -1 and 1 indicate negative correlation and positive correlation, respectively, and 0 represents independence. The PCC can be described using the formula:

$$r = \frac{\sum\limits_{i}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum\limits_{i}(x_i - \overline{x})^2}\sqrt{\sum\limits_{i}(y_i - \overline{y})^2}}$$

where (x_i, y_i) are pairs of data points, and $\overline{x}$ and $\overline{y}$ are the means of x and y, respectively. Using the PCC comes with several limitations in the context of this project:

- The PCC can only be used to identify linear relationships between variables.

- The data points of the two time series must be evenly distributed.

- The PPC does not address the issue that there may exist a time delay between one determinand's value and its effect on another determinand (and similarly for nearby sampling stations).

3.2 Cross correlation

Cross correlation is a generalisation of linear correlation, allowing us to account for possible time delays or "lags" between two series. It is a popular tool for analysing multiple time series and has been used for many decades on meteorological data; [3] uses cross correlation between precipitation patterns to forecast storms. More recently, cross correlation has been used to investigate relationships between weather and mosquito patterns [4] or weather and dengue fever [5].

In signal processing, cross correlation is a measure of the correlation between two signals as a function of the displacement of one signal to the other. In this calculation, one signal is fixed and the other is sliding. For this reason cross correlation is also commonly known as the "sliding dot product". **Figure 7** below shows the cross correlation $g * f$ of signals f and g. We observe that the cross correlation represents the overlap area of a fixed signal and a sliding signal.

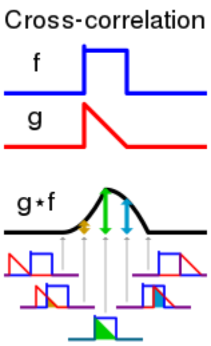


Figure 7: Cross correlation $g * f$ of two signals f and g [6]

The mathematical expression of cross correlation is

$$w(t) = f(t) \otimes g(t) = \Delta_{-\infty}^{\infty} f^*(\tau) g(\tau + t) d\tau$$

for continuous functions f and g and

$$w[t] = f[n] \otimes g[n] = f * [n] g[n + k]$$

for discrete f and g where f^*denotes the complex conjugate of f, and τ is the displacement.

[7] motivates the use of cross correlation on astronomical time series. The two time series which are studied are well-correlated but one is slightly time-shifted relative to the other.

In order to perform cross correlation, the data of the two time series must be evenly sampled. This issue can be dealt with by interpolating one of the two series between its available data points and using these interpolated points.[2] The resulting graph is shown in **Figure 8.**

[7] then calculates the **cross correlation function (CCF)**, calculating the cross correlation coefficient CCF(τ)

[2] This can be done using SciPy's interpolate function.[8]

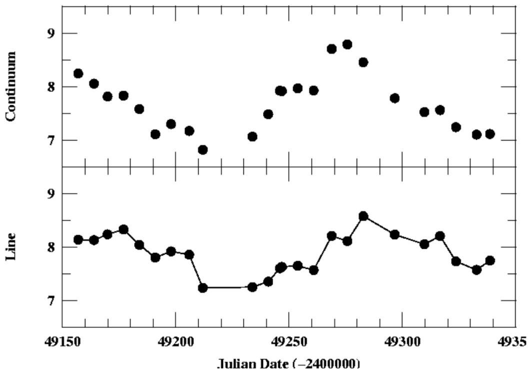


Figure 8: Continuum time series and interpolated line time series [7]

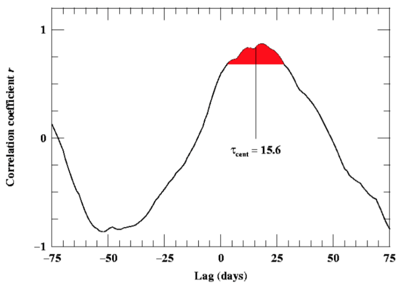


Figure 9: Cross correlation function of time series in Figure 8 [7]

for different values of lag. The CCF peaks at the value of τ which achieves maximal linear correlation coefficient for the two series. **Figure 9** shows the resulting cross correlation, indicating that the shift which achieves the optimal linear correlation is one of approximately 16 days.

The two series are then plotted again, with one of the series shifted according to the lag value. Performing linear correlation analysis on these two series then returns a correlation value which takes the lag into account. In short, the cross correlation function is the **"linear correlation coefficient as a function of time lag."** [9]

3.3 Discrete correlation

To achieve evenly sampled data, [7] interpolated one series with respect to the other. However, sometimes interpolation does not provide reliable results, particularly if there are too many large gaps in the time series, or if the data has been under sampled [7]. The **discrete correlation function (DCF)** makes less assumptions and can deal with irregularly sampled data. However, there is no straightforward implementation of this available, although a Python tool is available on Github [10]. For the purposes of this project, we use cross correlation but acknowledge that discrete correlation may be a more appropriate measure.

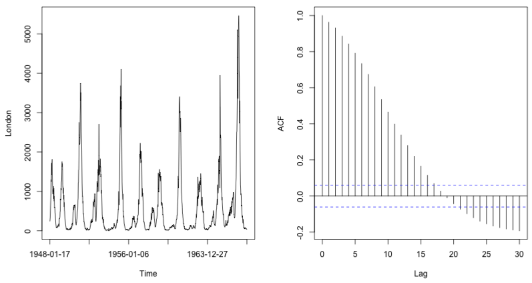


Figure 10: Measles auto correlation [11]

3.4 Autocorrelation

Auto correlation is a measure of the internal correlation within a time series which can be used to identify periodic behaviour. Note that the auto correlation at time lag 0 is always 1, because a series is perfectly correlated with itself. **Figure 10 left** shows a time series of measles incidents. This data is clearly periodic. **Figure 10 right** shows the auto correlation function of this time series. The auto correlation at time lag 1 is 0.96, indicating that the number of measles cases at a given week are similar to the number of measles the next week.

Therefore, to identify periodic behaviour, we identify the highest value the auto correlation function takes (besides 1 at 0) and determine whether that value indicates a strong correlation, that is, whether it has a magnitude close to 1. The lag at which this value occurs is the time resolution.

We therefore use auto correlation to investigate temporal correlations, where we want to determine whether the values of certain determinands are periodic in time. To do this, we pick one sampling station and find all of its samples for a particular determinand over a 5 year period and perform auto correlation analysis.

Note also that the cross correlation of a signal with itself is its auto correlation.

4 Models

4.1 Pollutant correlations

We restrict the measurements that we analyse to the top 20 determinands found in **Section 2.4**. We investigate the relationships between pairs of these determinands at sampling stations where both determinands are monitored. **Figure 11** presents the general workflow of the model. The model runs inside an iPython notebook and makes use of the numpy and pandas Python libraries. The output is a CSV file which contains a table of station IDs, years, time delay measurements, and correlation coefficient measurements.

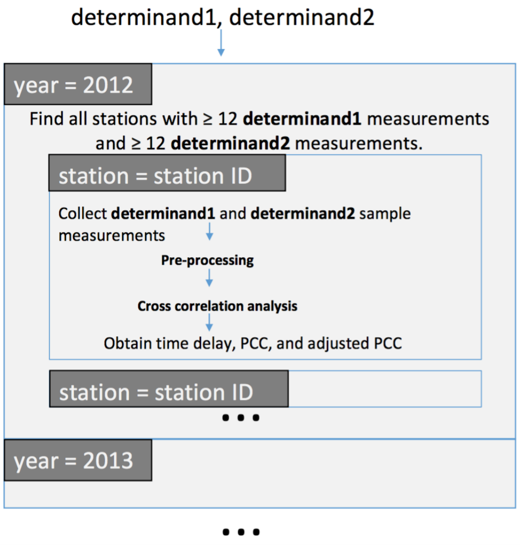

determinand1, determinand2
year = 2012
Find all stations with ≥ 12 **determinand1** measurements and ≥ 12 **determinand2** measurements.
station = station ID
Collect **determinand1** and **determinand2** sample measurements
Pre-processing
Cross correlation analysis
Obtain time delay, PCC, and adjusted PCC
station = station ID
• • •
year = 2013
• • •

Figure 11: Workflow of pollutant correlations model

We initially tried to investigate sampling stations that measured both determinands over a continuous 5-year period. However, there were very few of these, and their measurements for at least one of the years would often be infrequent. Therefore the model considers individual years.

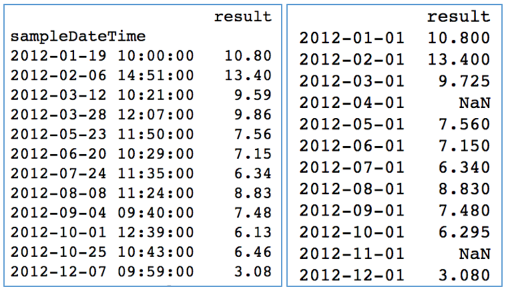
A lot of data pre-processing is involved to obtain correlations from the sparse data. Requiring ≥ 12 sample measurements from each sampling station did not necessarily mean that we had monthly measurements. In fact, even with this condition, there were often times gaps of over a month in the data. For each data set of sample measurements for a particular determinand from a particular station for a particular year, we:

1. Sort sample measurements by date

2. Reindex and resample the measurements to monthly measurements using the backfill method, that is, assigning the next value available in the time series

3. Discard the station if there are still missing values (other than December)

4. If the only missing value occurs in December, shorten the second time series to the January-November time period as well



```
                              result
sampleDateTime
2012-01-19 10:00:00    10.80
2012-02-06 14:51:00    13.40
2012-03-12 10:21:00     9.59
2012-03-28 12:07:00     9.86
2012-05-23 11:50:00     7.56
2012-06-20 10:29:00     7.15
2012-07-24 11:35:00     6.34
2012-08-08 11:24:00     8.83
2012-09-04 09:40:00     7.48
2012-10-01 12:39:00     6.13
2012-10-25 10:43:00     6.46
2012-12-07 09:59:00     3.08
```

```
            result
2012-01-01  10.800
2012-02-01  13.400
2012-03-01   9.725
2012-04-01     NaN
2012-05-01   7.560
2012-06-01   7.150
2012-07-01   6.340
2012-08-01   8.830
2012-09-01   7.480
2012-10-01   6.295
2012-11-01     NaN
2012-12-01   3.080
```

Figure 12: Raw data (left) and resampled and reindexed data (right)

Figure 12 shows an example of a time series before and after pre-processing Step 2. If there is a gap in the data for a month, a NaN ('not a number') is put in its position. If such a gap occurs before December, the time series is dropped. Reindexing and resampling in this way results in the loss of data and/or the creation of new data. While this is undesirable, these steps were necessary in order to have suitable input for correlation analysis as the sample measurements must be taken at equal time intervals.

Figure 13 shows two time series before reindexing and resampling; one is of nitrate as N measurements and the other of nitrogen as N. These time series are reindexed and resampled and the resulting cross correlation graph is used to obtain an indication of time delay. Using this time delay, we then adjust the time series and recalculate the PCC.

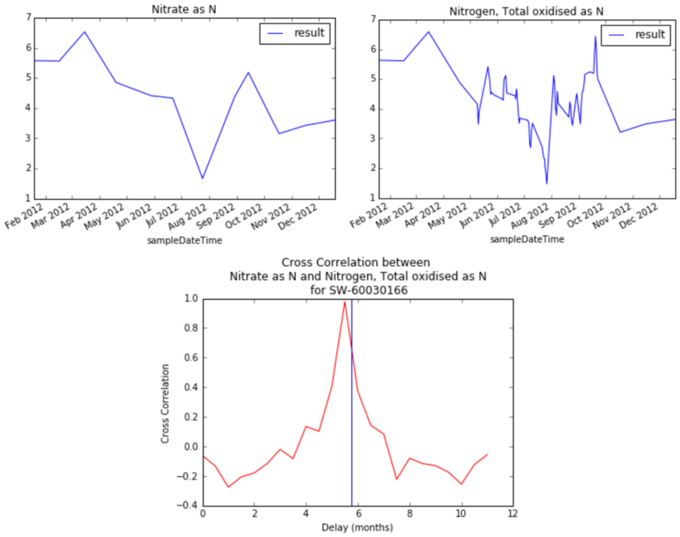


Figure 13: Time series (top) and resulting cross correlation graph (bottom)

We calculate the time delay and adjusted correlation coefficient in this way for the pairs of time series that remain. The model was applied to pairs of the top 20 determinands. This would result in 380 ordered pairs, corresponding to 160 determinand pairs. However, in fact only 290 ordered pairs were returned, corresponding to 145 determinand pairs. This is because if no non-null results are obtained, a CSV file is not created.[3]

Another program then loops through all of the CSV files, calculating for each CSV file the averages and standard deviations for time delay and adjusted correlation coefficient. We then conduct further processing to obtain only the statistically significant results. At this point the number of results drops greatly. This will be discussed further in **Section 5.**

4.2 Spatial correlations

For the spatial correlations model, we again only consider the top 20 determinands. **Figure 14** presents the general workflow of this model. We select a determinand and then find all stations that have made over $\geq$ 12 measurements of that determinand that year. We then iterate over each of these stations. Using the Water Quality API, we can then easily identify the sampling stations that are of 5 km distance from each other.

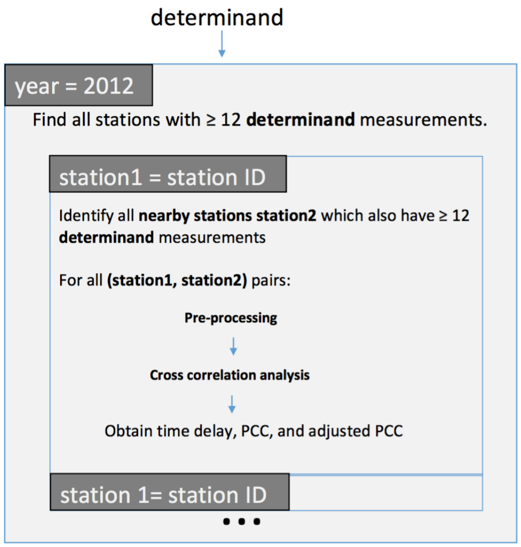

determinand
year = 2012
Find all stations with ≥ 12 **determinand** measurements.
station1 = station ID
Identify all **nearby stations station2** which also have ≥ 12 **determinand** measurements
For all **(station1, station2)** pairs:
Pre-processing
Cross correlation analysis
Obtain time delay, PCC, and adjusted PCC
station 1= station ID
• • •

Figure 14: Workflow of spatial correlations model

For each pair of stations that we find, we perform similar pre-processing steps as before. We repeat this for each of the top 20 determinands. The post-processing steps are nearly identical to those described for the pollutant correlation model.

4.3 Temporal correlations

To identify temporal correlations, we first look for sampling stations which have made many measurements of our chosen determinand. Rather than reindexing and resampling the measurements to monthly measurements, we now resample to weekly measurements. This is necessary as we require finer time resolutions in order to identify periodic behaviour.

We would then take the time series which do not have any gaps and perform auto correlation analysis. Identifying the position at which the maximum (but not equal to 1 or 0) auto correlation value occurs would indicate a time lag. An auto correlation value of magnitude close to 1 would suggest that the time lag found represents the time resolution of the determinand's periodicity.

However, if we make the requirement to drop any time series which have gaps of over a week, as we must do in order to have equally fine spaced time intervals for auto correlation analysis, we remain with no time series. We therefore find that the sparsity of the data is a severely limiting factor in our temporal correlation analysis.

One potential solution is to collect samples from a region of sampling stations instead. However, this could introduce discrepancies in our data and would only be justifiable if we could show that the sampling stations we collect samples from are all closely related and generally record similar measurement values.

5 Results

Having developed iPython notebooks for each of our models and using the Water Quality API, we make the following findings.

5.1 Pollutant correlations

The output of our pollutant correlations model was a collection of CSV files for each determinand pair which includes time delay and adjusted correlation coefficient measurements. We then looped through these CSV files, calculating the averages and standard deviations for time delay and adjusted correlation coefficient for every determinand pair. From this, we limit ourselves to results with standard deviations of less than 1. The number of results drops dramatically from 145 determinand pairs to only 28 pairs.

To identify the determinand pairs which exhibit positive correlation, we look at the results with an average correlation coefficient value above 0.9. These results are presented in **Table 1.** None of these results are surprising. The first and third pairs are clearly measuring nearly the same property. The second pair is also unsurprising, as nitrate is a compound of nitrogen and oxygen. Although these results may not lead to any new insights from an environmental point of view, the fact that they unsurprising serves to verify the pollutant correlation model created.

5.2 Spatial correlations

Again, we limit ourselves to results with standard deviations of less than 1. This results in a drop in the number of results from 20 to 4. When we further require a correlation coefficient of magnitude above 0.9, we obtain one result, presented in **Table 2.** This result suggests that the measure of the temperature of the water correlates between nearby

[3] If the cross correlation is equal to 'NaN', the result is dropped.

Determinand 1	Determinand 2	Time Delay	Correlation Coefficient
Oxygen, Dissolved as O2	Oxygen, Dissolved, % Saturation	-0.25	0.916053
Nitrogen, Total Oxidised as N	Nitrate as N	-0.25	0.996182
Enterococci: Intestinal: Confirmed	Enterococci: Intestinal: Presumptive	-0.25	0.998667

Table 1: Positively correlated determinands

sampling stations, that is, that the water temperature at one station can probably be used to infer the temperature at another station.

Determinand	Time Delay	Correlation
Temperature Water	-0.25	0.970625

Table 2: Positively correlated spatial correlation

5.3 Temporal correlations

We found the sparsity of the data to be severely limiting, and because of this we were unable to find any results.

6 Conclusions

In this section we evaluate our findings, suggest improvements, and give recommendations on what could be done moving forwards.

6.1 Discussion

The major obstacle in all three of our correlation models was the sparsity of the data. Reindexing and resampling each time series to monthly (or weekly) measurements was done in order to obtain data suitable for cross correlation and auto correlation analysis. However, often the data was not frequent enough for the resampling to provide suitable data, for instance, with no month-long gaps. Even if the data was suitable, one must remember that these methods involve removing or adding new data.

The unsurprising results in the pollutant correlation model should be taken as a positive indication suggesting that other meaningful (and perhaps less unsurprising) correlations could also be found with more data. When looking for correlations there is always a danger of coming across "spurious correlations" so it is at least reassuring to see that the correlations identified make sense. We believe that the three models described in **Section 4** are a good start at identifying correlations and could be used to yield meaningful results.

6.2 Improvements

Reindexing and resampling each time series was done with a 'one size fits all' approach. In the case of the pollutant correlation and the spatial correlation models, this meant taking the stations with $\geq$12 measurements and reindexing and resampling time series data to monthly measurements. This method was chosen in order to pre-process the data efficiently on a large data set. It is possible that there may be better ways to resample the time series if we consider each pair of time series on a case-by-case basis. If, for instance, both series provide frequent sample measurements for September and October, we could look at the correlation between the two series over this shorter time frame.

In **Section 3,** we gave an overview of the cross correlation and auto correlation measures used, but also suggested that the discrete correlation function may be a more appropriate choice for our data. We believe that this alternative measure may have allowed us to avoid some of the issues with the data that we had.

In terms of spatial correlation, one existing limitation of our model is that it does not consider the purposes of the various sampling stations. In some initial data analysis not shown, we found that often pairs of nearby sampling stations consisted of one sewage disposal next to a sewage treatment work. Unsurprisingly, such two sampling stations exhibit vastly different sample measurements. We therefore believe that our results could have been improved if we also took into account the purpose of the sampling station and treated different sampling stations differently.

In our work, we have limited ourselves to the 'Wessex' area. However, we could have considered other areas as well as it is possible that sampling stations in other areas are being measured more frequently. We also could have looked at the bathing water quality data, where sampling stations provide more regular measurements over the May-October period. However, the scope of our project would have then been very different.

6.3 Future Work

During our talks with the EA, an area of interest for them was the comparison of their data to similar data collected by some local councils. The EA wanted to know if they could work with local councils to reduce monitoring. For instance, the Bristol City Council collects its own river water quality data and makes it available through the data.gov.uk site. This dataset is in a different format to that of the EA. Most notably, sampling station locations are indicated by site references. Another file called "River site National Grid references" then provides the NGR (National Grid Reference) for each site reference. Because there are only several site references, we manually converted each NGR into latitude and longitude coordinates using an online tool.

Figure 15 presents a map of the Bristol City Council and the EA sampling stations. The EA sampling stations are indicated by orange and grey dots, with orange dots indicat-

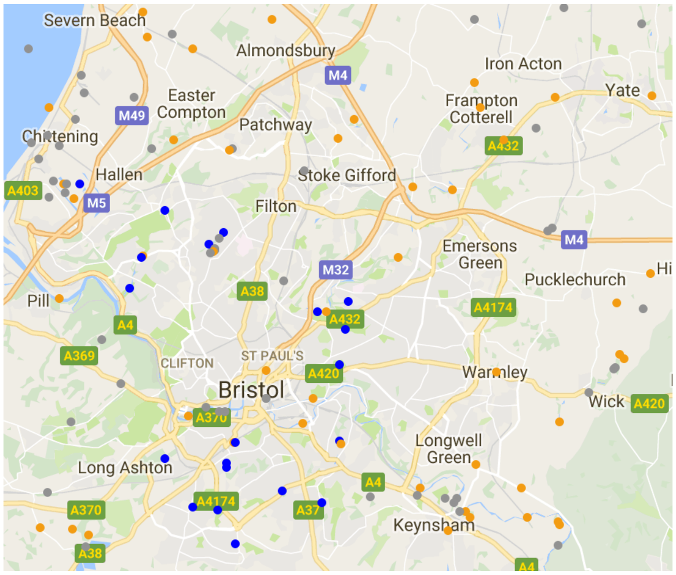


Figure 15: Map of EA (orange, grey) and Bristol City Council (blue) sampling stations

ing sampling stations active in 2016 and grey dots indicating sampling stations that inactive in 2016. The blue dots indicate Bristol City Council sampling stations. From the map we can identify several areas where both the EA and Bristol City Council are making measurements, such as Brislington brook, but also several areas where the Bristol City Council is making measurements but the EA is not.

While the Bristol City Council makes measurements on only a small handful of determinands, and makes them too infrequently for correlation analysis, it may nevertheless be useful for the EA to know what data the Bristol City Council is collecting to see how they could benefit from working together. Further work could be done by investigating, for instance, whether the nearby sampling stations are providing results for the same dates.

Returning back to our correlation models, if we had more time we would have liked to have investigated the discrete correlation function, which may have allowed us to overcome some of the issues in our data. Furthermore, combining data from all areas of England and not just Wessex may have allowed us to identify more correlations.

Some of the initial work for this project involved finding tools to plot the sampling stations on a map with labels. The Google Maps API could be used to this end, although we used a special iPython notebook plugin for the Google Maps API. Given more time, this project could be extended by developing an online tool which could then allow the EA to explore various correlations within an interactive map.

Acknowledgments

We would like to thank John Burke and the rest of the staff that we had the opportunity to meet with at the Environmental Agency.

References

[1] Water quality data archive. Accessed: 2017-04-01.

[2] Pascal Bugnion. gmaps: Google maps for jupyter notebooks.

[3] GL Austin and A Bellon. The use of digital weather radar records for short-term precipitation forecasting. *Quarterly Journal of the Royal Meteorological Society*, 100(426):658–664, 1974.

[4] Ting-Wu Chuang, Edward L Ionides, Randall G Knepper, William W Stanuszek, Edward D Walker, and Mark L Wilson. Cross-correlation map analyses show weather variation influences on mosquito abundance patterns in saginaw county, michigan, 1989–2005. *Journal of medical entomology*, 49(4):851–858, 2012.

[5] Pei-Chih Wu, How-Ran Guo, Shih-Chun Lung, Chuan-Yao Lin, and Huey-Jen Su. Weather as an effective predictor for occurrence of dengue fever in taiwan. *Acta tropica*, 103(1):50–57, 2007.

[6] Wikipedia. Cross-correlation. Accessed: 2017-04-11.

[7] Bradley M Peterson. Variability of active galactic nuclei. *The Starburst–AGN Connection*, pages 3–67, 2001.

[8] numpy.correlate. Accessed: 2017-04-11.

[9] Bradley M. Peterson. Time series analysis in studies of agn variability. The Ohio State University.

[10] Damien Robertson. A python cross correlation command line tool for unevenly sampled time series.

[11] Guy Nason. Time series introduction.