Analysing relationship between water quality and land use

Gizat Makhanov
tq18441@bristol.ac.uk

Ruben Powar
wz18202@bristol.ac.uk

Durgadevi Rajendran
lx18327@bristol.ac.uk

Jeonghyun Kim
cc18316@bristol.ac.uk

Jinyang Zhan
qi18235@bristol.ac.uk

ABSTRACT

This study analyses the correlation between quality of water in reservoirs in England and the density of crop field. Two types of data were used: 1) data from various water sources in England were collected since 2000, and 2) density map of crop fields in England for 2010. The analysis shows a clear correlation between the number of crop fields in an area and the nitrate levels in nearby water reservoirs.

The study also includes the main tasks and tools used to complete this analysis available on Github:
github.com/gizat/ads-ea

INTRODUCTION

An increase in the practice of inorganic agriculture in the recent years had significantly affected the water quality in natural resources in the United Kingdom [3][5]. The main source of pollutants is the usage of fertilisers. The chemicals in fertilisers are absorbed by soil and spread to nearby water resources through surface runoffs or underground water channels. Surface runoff in agricultural areas is additionally enriched with different chemicals, nutrients, and sediments. Moreover, different types and coverage rates of vegetative surfaces modify the land surface characteristics, water balance, hydro-logic cycle, and the surface water temperature through various natural phenomena known as evapotranspiration, interception, percolation and absorption. The polluted quality of water in runoffs, streams and ground water continue polluting the receiving water resources [10].

With the decreasing trend in organic farming and the expanding plethora of different types of fertilisers manufactured is seen alongside a diminishing awareness on the ill-effects of fertilisers and increasing practice of inorganic farming. The effluents from agricultural land usage where inorganic farming is practiced, vary with the types of fertilisers used. Inorganic fertilisers are made up of a number of chemicals, with some acting as contaminants when they pollute the water by inorganic farming. Nitrogen is one of the primary chemicals essential for plants and crops and fertilisers contain constituents which has nitrogen. Use of nitrogen-based fertilisers in excess of permissible levels, can cause damage to the soil where farming is practiced and consequently cause contamination of groundwater.

The existing methods and processes to treat the contaminated water and remove the nitrates present are very expensive, highly complex and have some constraints that makes it difficult for governing bodies of rural areas to invest in such removal methods. This is one of the reasons why there are cases of communities migrating to other locations in search of clean water without the contaminants. [9].

The presence of nitrates in untreated groundwater and consequently drinking water can cause serious health issues to humans, especially woman and infants over a long term [4]. With a growing awareness of the detrimental effects on health and well-being of humans due to such determinants in water, there are a number of environmental agencies including government agencies that are studying the effects of harmful effluents on water quality. The chemicals present in water can be measured in multiple ways and certain ones can be used as an indicator of water quality. However, many of these might be highly impractical or expensive to measure and may not always lead to correct measurements or results. Water quality standards depend not just on what is in the water but also what the water is used for (e.g. drinking or watering crops). [6]. Such standards are a key consideration factor for water quality measurements as it helps in focussing the studies on a smaller and fine-tuned set of quantifiable determinands. According to the Summary of the Chief Inspector's report for drinking water in England [7] The maximum acceptable level for Nitrates in drinking water is 50mg/L, with 10mg/L serving as a goal maximum value according to the EPA [2].

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. The Environment Agency (EA), acting under the Department for Environment, Food & Rural Affairs (DEFRA) of the government of UK, collects millions of measurements on water quality around England every year. This agency collected over 14 million water quality measurements made of over 200 different determinands in 2016.

Our first stage of analysis, for this project, was based on the water quality provided by EA since 2000. Further data related to crops were collected from Rural payments agency. With a careful and methodical application of data analysis concepts on the available datasets, the final outcome of this project supported our intuition that there was a high correlation between land usage and components of water quality. Following sections of this document will provide details on the approach, methodology, analysis and findings of the research.

COLLECTION, EXPLORATION, AND PRE-PROCESSING

The main data provided to us for this project was collected by the environmental agency (EA) [1], measuring roughly 1100 differing forms of contaminants in water. The EA is a public body sponsored by the department of environment whose responsibility relates to the protection of the environment in England, hence that type of data collected and the geographical range being limited to England rather than the entirety of the UK. The data was segmented into individual CSV files each relating to a year spanning the period 2000 to the first quarter of 2019. On average each annual CSV is roughly 1Gb in size, containing 3,000,000 values.

Many attributes of the data, while useful for understanding the nature of the information would serve no practical function in the context of this project, for example the following attributes all represent the same information on the geographical location of each sample: id, sample.SamplingPoint, sample.SamplingPoint.notation, sample.SamplingPoint.label. This facilitates a reduction in the dimensionality of the data while losing no actual information, allowing the code produced to run on data of a much reduced size. The contaminant type is stored in an attribute result, contaminant type in determinand.label, determinand.definition, and with time-stamps given in sample.sampleDateTime.

The data for a single year was further partitioned into a variety of working test sets: collection of heavy metals, and a collection of nitrates and nitrites sampled in multiple different forms. The intention behind this subsetting was to provide smaller files on which to work, thus potentially providing some early confirmation for theories patterns without having to work on the entirety of the dataset, which would be too computationally demanding to be done in a timely manner. The simple year selected was 2010, as this is the year to which the DEFRA map (discussed below) pertains. After this removal of certain irrelevant attributes and further partitioning the data was prepared sufficiently to be used for spatial analysis.

This subsetting was also carried out on the entire data set in order to facilitate some temporal analysis of the data, i.e. trends in single locations over a given time period.

Our initial approach for the research had plans to assess the impact of industries on water quality along with agriculture. We relied on the industries data from www.industryabout.com for this analysis. This website had a varied set of datasets related to a number of fields - one of the key datasets related to industries had location coordinates for about 2100 industries spanning across 51 sectors ranging from Aerospace to Oil refining. Whilst this data seemed to cover most of the industries in UK, there was a sparsity of corresponding data for water quality and due to time consrtraints, we limited our scope to assessing the impact on water quality due to agriculture.

In order to analyse the relationship between water quality and land usage, a dataset showing crop field areas was needed. There are several free and commercial sources that provide this kind of dataset. After analysing all options, dataset from the Department for Environment, Food & Rural Affairs (DEFRA) was chosen.

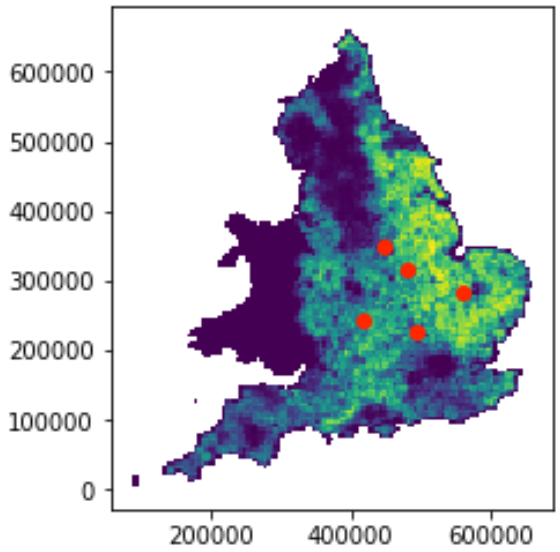

Figure 1. Map of crop areas in England and random sampling points with water quality data, 2010

DEFRA is a UK government department that provides detailed annual statistics and maps on the structure of the agricultural industry in England and the UK. Their datasets show land and crop areas, agricultural trends, livestock populations and agricultural workforce estimates for England and UK. The results come from long-running surveys conducted every summer. Their other dataset is generated using supervised classification from Sentinel-1 and Sentinel-2 images.

Survey approach was used for this study. The main limitation of this dataset is the estimations made for the absent non-respondents, which comprised about 30%. Estimates are also made of the exact locations of land areas.

The website didn't provide the data in the format needed. Upon contacting and explaining the project, the DERFA office shared data directly with us in the form of shapefiles. The dataset had columns for density indices of different types of crops. Each crop density index was associated with a rectangular polygon shape represented by four geographic locations in the easting and northing standard. Plotting all cells generated a colour map of England with various density levels for a specific crop. A map with all crop types is shown in Figure 1.

The first operation was to translate the rectangular polygon coordinates into the latitude and longitude standard. Many approaches exist, but GeoPandas' *to_crs* function was chosen to change the projection.

It was possible to view the sampling points on a crop map without any integration. However, further analysis would be impossible without linking these two datasets. After studying the datasets, association of sampling points with rectangular cells using geographic coordinates was chosen to be the more effective approach.

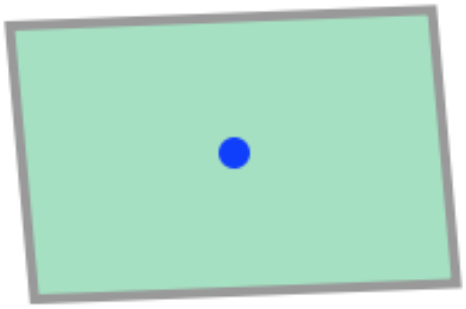

Figure 2. A rectangular shaped cell and it's centroid

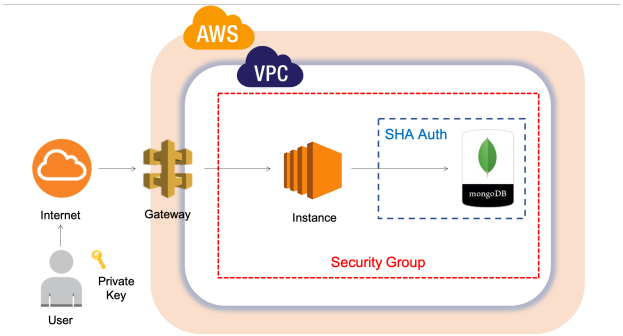

Figure 3. Database architecture

Association of a sampling point to a rectangular cell was found by calculating the Euclidean distance between each sampling point and the centroids of each rectangular cell. Polygon shapes were converted to geographical points. The shortest distance from a sampling point and a geographical point representing a rectangular cell would be chosen (Figure 2).

As there were 6585 centroids and 27,797 sampling points, it would take approximately 183 million operations to find the Euclidean distances. A more efficient approach was to use vectorisation and broadcasting. The approach took 35 seconds to find all Euclidean distances.

DATABASE AND SECURITY

Platform

Cloud platform was chosen as the main storage for the datasets collected. Storing database in a cloud had multiple advantages over storing data locally:

- **Collaboration**. Three team members were working with the database directly and it was convenient to access one version of the database on a cloud, rather than storing three copies on local computers.
- **Security**. All cloud platforms provide high level of security of client data using firewalls, security protocols, and physical restriction to storage areas. Also, while data on local computers could be lost due to equipment loss or malfunctioning, cloud providers guaranteed the client data would be secure.
- **Flexibility**. Cloud platforms allow for easy expansion of storage sizes.

Three popular cloud platforms were assessed: Amazon AWS, Google Cloud Platform, and Microsoft Azure. Each of these providers went through the following two evaluation criteria:

- **Performance**. AWS provided highest memory size and CPU clocks. Google Cloud Platform had higher networking speed. Microsoft Azure performed worse than the other two.
- **Price**. All three provide initial credits of similar value valid for 12 months.

Based on this evalution, Amazon AWS was chosen as the cloud platform.

Database

Although the team members had prior experience with relational databases, like SQL, document databases or NoSQL, like MongoDB, were considered. MongoDB performs better with inserts, updates and other simple queries [8]. As the data set size used in this study was large, 20 GB, it was important to deploy a scalable database management system. NoSQL can deal with scalability issues with the data shard mechanism. Price was another criteria, which means open-source available database is appropriate to this project. Last consideration is flexibility. Even if design of data model is changed, it should be easy to modify and manage the new data model.

MongoDB was deployed on Amazon AWS based on the above evaluation. Figure 3 shows overall database infrastructure. To access an instance in AWS, an approved key pair is required. Although a user with the key pair can pass through a gateway to enter Amazon Virtual Private Cloud (VPC), IP address should be included in a security group that is regulated by an owner. Lastly, a packet from a user is verified by MongoDB with SHA hashing algorithm whether it corresponds to its ID and password.

Security

Although security is not a key issue regarding the handling of our data, It is essential to handle security issues in this project due to potential confidentiality or integrity concerns should the system be implemented in industry. Our first security layer is provided by cloud. Cloud has their firewall, which is prevent to malicious access to their private network, near the gateway. Inside of the private network, in addition, we designed one more security layer for safety of database. We employed one of authentication mechanism provided by MongoDB. We chose SCRAM SHA-256 hashing algorithm. There are more options like Kerberos and X.509. However, Kerberos is available for subscribers and X.509 needs three replica servers to trust one server, which means it is hard to implement on free tier. Therefore, we choose SCRAM authentication mechanism for our secondary security layer.

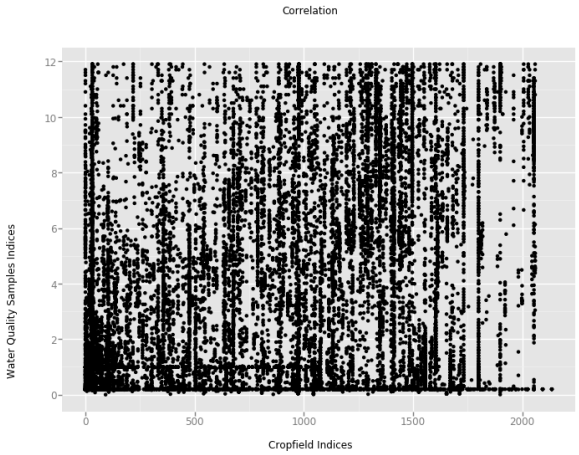

Figure 4. Scatterplot with all samples before binning

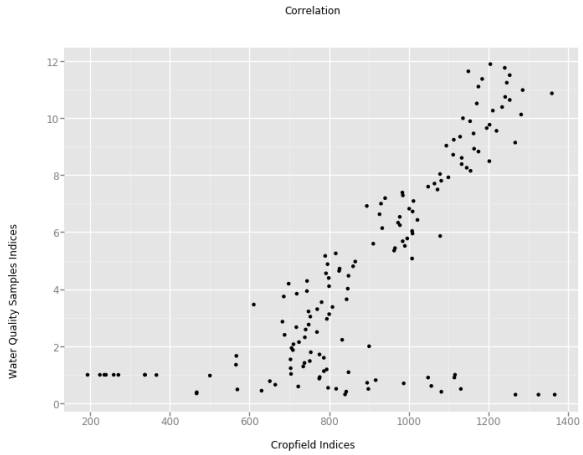

Figure 5. Scatterplot of samples after binning

DATA ANALYSIS

The analysis of the data was separated into two key sections: spatial analysis (pertaining to the trends between different sampling locations in relation to geographical factors with time invariant) and temporal analysis (pertaining to variation of single locations over a sampled period of time). These analyses give intuition into some of the driving factors behind the contaminant variation and allow for exploration of trends over time, both seasonally and over a longer period.

Spatial analysis

For this portion of the analysis the data from year 2010 was used as it corresponds with the date of origin of the cropfield map obtained.

As is apparent in the visualisation tool produced (Figure 12) there is some correlation between areas of high cropland density and high nitrate levels in the water samples. Some quantification was required however, in order to corroborate the apparent findings of the visualisation. When plotting all samples from the dataset, with the cropfield density metric on the y-axis, and the actual nitrate level of the sample on the x-axis the graph in Figure 4 was produced.

This shows how noisy the data is. There appears to be a significant amount of variation in nitrate level detected between sites of similar usage. This did not corroborate the hypothesis that there existed some fairly evident correlation as observed in the webtool. With a Pearson's correlation coefficient of 0.29 we can see there is some very weak linear correlation. There are however a few key areas of note that imply some underlying logic not entirely represented by the correlation coefficient. Looking at the origin there appears to be a high density of points, supporting the notion that at areas of low cropfield index there is an expected low nitrate level observed. Also, at the extremes with low cropfield index and high observed nitrate levels (and vice versa) there is a relative sparsity in samples. This further supports that there would be some form of correlation between high nitrate level and heavily agricultural areas.

In order to make further sense of the trends and attempt to extract the apparent underlying information the data was sorted by nitrate levels and binned accordingly. This gave a reduction in the amount of noise in the data as in each bin the nitrate level and cropfield index were averaged. With bins of size 100 the following graph was produced:

As is evident in Figure 5 the intuition resulting from observing the webtool is supported, with a Pearson's correlation coefficient of 0.71. This shows a strong positive linear correlation between cropfield index and nitrate levels: the more heavily agricultural a location is the higher the observed nitrate levels observed, implying some causational relationship between agricultural practices such as fertilisation and water quality.

Temporal analysis

In order to carry out temporal analysis on the data a small amount of further preprocessing was required. The analysis is carried out on individual locations, observing how the observed measurements vary over time. This therefore required all observations to be extracted for the selected location from the entire time period sampled.

Once the data was prepared it was then plotted, with nitrate level on the y-axis and time on the x-axis. Two key types of location were analysed: 1. protected areas with little to no agricultural activity 2. heavily agricultural areas with high cropfield index.

Lincolnshire Sampling Point

This location is fairly central to the county of Lincolnshire, which is responsible for 1/8 of the UK's food, including the majority of the country's vegetables. This highly level of agriculture is reflected in the the cropfield index at this location of 13.5, among the highest of any location. This location therefore serves as a good example of the highly agricultural extreme of land usage. Figure 6 below shows all the data for this location available, between 2010 and 2017 with a linear model fitted to the data.

There is considerably variation in observed Nitrate values, appearing to be random noise. However, when fitted with

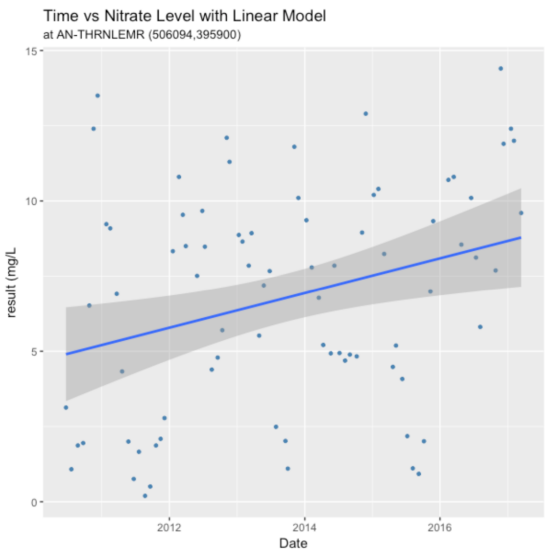

Figure 6. Local regression using LOESS method of time vs nitrate level

a non-linear model a pattern begins to emerge. The model selected to analyse the non linear trend in the data was (locally estimated scatterplot smoothing), a generalisation of moving average and polynomial regression. There appears to be a fairly visible seasonal trend, with a period of one year.

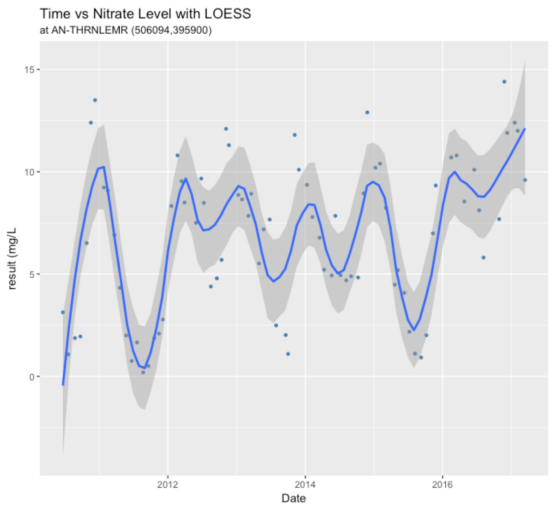

Figure 7. Linear model of time vs nitrate level

Around the beginning of each year the observed Nitrate levels appear to peak, then dropping considerably reaching a minimum just after the midpoint of the year. This corresponds with agricultural practices for the location, with vegetables typically sown in February to early March and recommended fertilising occurring roughly 6 weeks before this. At first glance this may seem as though the model has overfit to the data, however this pattern is evident in the vast majority of locations of high crop-field index. This, coupled with the regularity of the fluctuation

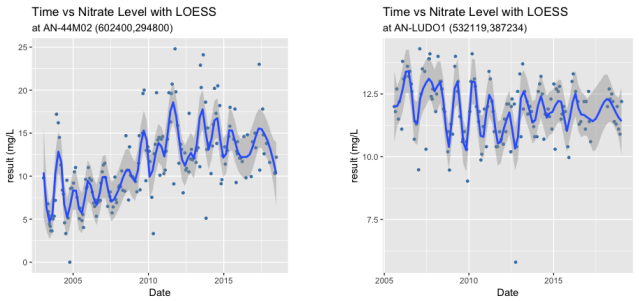

Figure 8. Two more agricultural locations showing high degree of seasonality

and the fact that it coincides with agricultural practices in the region is strong evidence to suggest that this patter is a result of human interaction with the farmland.

West Yorkshire Sampling Point

This location has a very low crop field index = 0.004, due to being in a protected area. This makes it a good example of the other extreme: a completely uncultivated location.

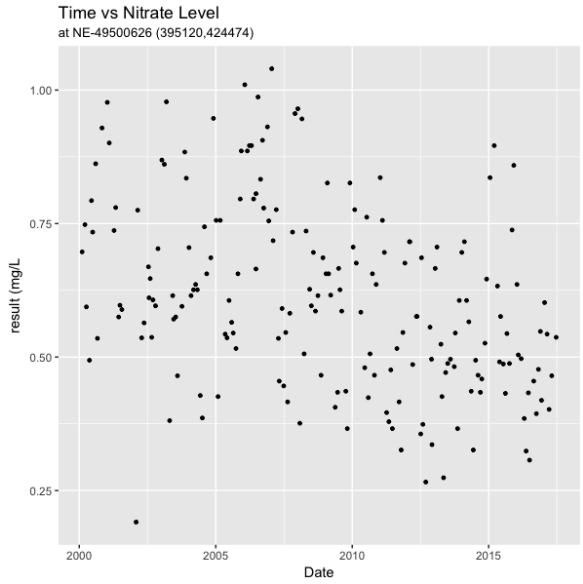

Figure 9. Scatter plot of observations on non-agricultural area

The readings here appear to be far less coherent, not displaying any of the seasonality of a highly agricultural area. Upon further analysis the data appears to be normally distributed, as can be seen in the graphs in Figure 9 the Quantile-Quantile plot and in Figure 10 probability-probability plot (bottom right) both show the empirical data to closely lie in correlation with the theoretical distribution, showing that the noise observed in the variance of the observed nitrates levels to be normally distributed.

Holt-Winters Modelling

In order to produce some predictive models of the data factoring in periodic seasonality Holt-Winters filtering [11] was

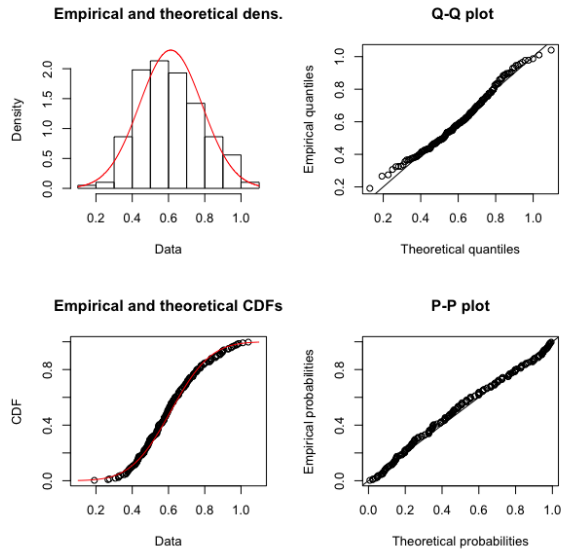

Figure 10. Analysis of distribution

applied, a popular seasonal forecasting technique. Holt winters is a form of exponential smoothing, using exponentially moving averages and factoring in seasonality. This method requires a complete time series of the data, whereas the data from the EA is more sparse. Therefore the missing data was required to be imputed. The first methods attempted for this were simply replacing all missing values by 0 and replacing missing values by the mean of the present data. These naive attempts to impute data were not successful however due to the high degree of sparseness. Had the data been more complete these methods still would likely not have been sufficient as imputation using a fixed value would interfere with the seasonality of the data. Linear interpolation was attempted to impute the data. This essentially fills in missing data along a linear line between existing data values. This preserved many of the qualities of seasonality observed, producing much more stable Holt-Winters smoothing than the previous attempts.

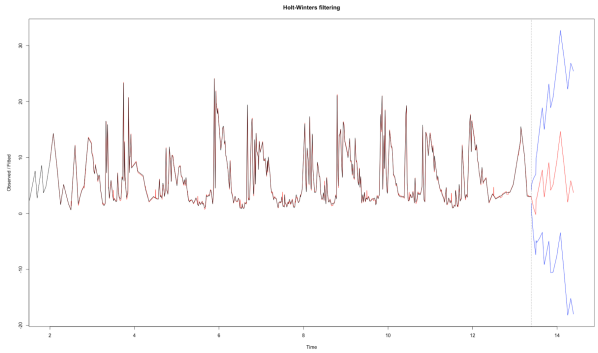

Figure 11. Holt Winters model and prediction

The x-axis on this graph represents time in years, with evident periodicity, and the y-axis shows actual and fitted Nitrate levels in mg/L. The black line represents the actual observed data after having been imputed using linear interpolation. The red line running up to the vertical dashed line represents the Holt-Winters smoothing of the data, predicting the trend for the next period after the dashed line. This prediction maintains the seasonality represented in the observed data, and is able to outperfrom and of the previous regressive techniques implemented, i.e. linear modelling, polynomial regression, and local polynomial regression.

Key points to note when applying Holt-Winters forecasting on our collected data are the following: At certain sampling points, although obvious periodicity is evident it does not span the entirety of the sampled period. This may be due to land owners allowing land to go fallow in order for the health of the soil to improve so as to not deplete it. This may result in a year or more of inactivity, and therefore no fertilisation, effecting any model formed, causing a larger amount of uncertainty. This is evident at the beginning of Figure 11. In order to produce a predictive model that is as reliable as possible Holt-Winters should only be applied to active periods of farming for a given area. This prediction of when farmers may decided to allow land to go fallow is beyond the scope of the analyses produced in this project.

DATA VISUALISATION

Due to the geographical structure of both the water quality and the crop area datasets, mapping was the most effective communication tool to visualise the data. There are many providers of online maps, but the evaluation was conducted among the most popular three: Google Maps, Mapbox, OpenStreetMap, etc. Studying the supporting toolset of each of these online maps, Mapbox was chosen to be the most comprehensive and aesthetically better than the others.

Mapbox supports a wide range of file types: CSV, GeoJSON, Shapefile, etc. Connection to the API can be accomplished in a JavaScript file.

The first step was to add a layer of the crop area grids on top of England on the Mapbox map. Choropleth map styling was used to differentiate levels of density of crop areas. As such, darker green colors represent areas with more dense crop fields, and lighter green – less dense (Figure 12 (left)).

This mapping shows visual correlations between the level of nitrates in sampling points and crop area density.

The next step was to load all the sampling points on the map, differentiated by colour. The colours are associated with the nitrate level at each sampling point. Here, the dark green circles represent sampling points with nitrate levels above 10 mg/L, while the lighter shades of purple – 0 mg/L (Figure 12 (middle)). These levels were obtained from the EPA and DOI as previously stated.

Another tool used to visualise data was the ggplot2 package. It is a visualisation package for the statistical programming language R. It was chosen over other visualisation packages for being aesthetically clean and ease of use. All the graphs shown in the Data Analysis section were created using ggplot2.

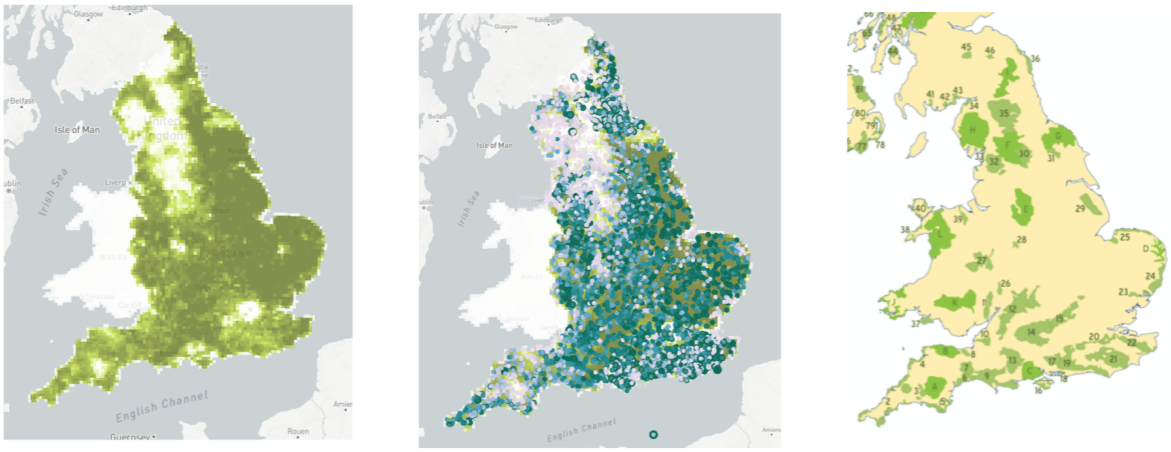

Figure 12. (left) Map of crop areas in England, 2010; (middle) Map of crop areas in England and all sampling points coloured by the level of nitrates found in samples, 2010; (right) Map of national parks in England.

CONCLUSION

This project sought to analyse the effects of land usage on water quality in England based on given data from the Environment Agency and collected data from various other governmental bodies. The specific case study selected was the use of chemicals present in fertilisers being used by the agricultural industry, a large component of land use in England. The project focused on Nitrates and Nitrites, compounds ubiquitous with fertilisers for vegetable and wheat crops.

Tools for analysis and visualisation were developed, allowing a user to see correlations between areas of high agricultural usage and high observed Nitrate levels in water samples. This was then further quantitatively shown through the analytical tools developed in python and R. A clear correlation between land usage and nitrate levels was evident, with a Pearson's correlation coefficient of >0.7-indicating strong a strong linear relationship-on the cleaned data. On a more local scale when observing individual location measurements over time it was demonstrated that areas entirely unused for agricultural purposes not only were nitrate levels typically consistently low, but also showed no tangible seasonality or variation over time. On the other hand, heavily agricultural areas showed a definite seasonality that was correlatory to farming techniques.

Developing on these findings, predictive models utilising the features of seasonality and general trends were developed, showing a reasonable degree of accuracy over less sophisticated methods. These predictive models show promise in further refinement with clear future optimisations.

The work done in this project could further be extended to include a wider range or chemical contaminants analysed. While the data focused on in this report only referred to Nitrates, the visualisation tools and basic analytical tools were developed for Nitrates and Nitrites. Data was also collected on location of industrial objects as well as heavy metal contaminants from the Environment Agency dataset. This would likely be the next area to analyse in the further implementation of these tools. The predictive modelling, while showing promise, could certainly be optimised further in order to bring it to a level appropriate for any potential use in industry. Potential optimisations would include further development of pre-processing and aggregations of multiple models from adjacent geographical areas to produce a more robust model and predictions.

Application in industry could include:

- Private use by land owners
- Governemental use

These tools could allow private land owners or farmers to monitor their own effects on their environment in order to ensure they self regulate and remain responsible in their farming practices, avoiding any potential repercussions. It may also be useful as a tool to help optimise farming practices. When combined with other data such as recorded crop yields and other environmental data such as weather patterns, farmers may be able to optimise their planting and fertilising pattern in order to maximise potential crop yields, and therefore increase profitability.

Governmental agencies may also find these tools useful in the enforcement and regulation of chemical runoff and pollution. As this data is already widely collected by governmental agencies these tools could simple be applied to this data to predict possible future areas of concern, allowing for streamlined regulatory focus, as well as identifying areas or specific farmers who frequently approach or exceed the upper bound of acceptable observed pollutant levels in water samples. In theory these tools could be applied to any geographical data and are not restricted just to the UK or England.

REFERENCES

1. 2019. Environment Agency - Water Quality Archive. https://environment.data.gov.uk/water-quality/view/download/new. (2019). Accessed: 2019-03-20.
2. Environmental Protection Agency. 2014. EPA - Nitrate and Nitrites. https://www.epa.gov/sites/production/files/2014-05/documents/nitrates_nitrites_presentation.pdf. (2014). Accessed: 2019-03-28.
3. RE Brazier, GS Bilotta, and PM Haygarth. 2007. A perspective on the role of lowland, agricultural grasslands in contributing to erosion and water quality problems in the UK. *Earth Surface Processes and Landforms: The Journal of the British Geomorphological Research Group* 32, 6 (2007), 964–967.
4. Jianyao Chen, Changyuan Tang, Yasuo Sakura, Jingjie Yu, and Yoshihiro Fukushima. 2005. Nitrate pollution from agriculture in different hydrogeological zones of the regional groundwater flow system in the North China Plain. *Hydrogeology Journal* 13, 3 (2005), 481–492.
5. A Louise Heathwaite and PJ Johnes. 1996. Contribution of nitrogen species and phosphorus fractions to stream water quality in agricultural catchments. *Hydrological processes* 10, 7 (1996), 971–983.
6. Joseph Holden, Philip M Haygarth, Nicola Dunn, Jim Harris, Robert C Harris, Ann Humble, Alan Jenkins, Jannette MacDonald, Dan F McGonigle, Theresa Meacham, and others. 2017. Water quality and UK agriculture: challenges and opportunities. *Wiley Interdisciplinary Reviews: Water* 4, 2 (2017), e1201.
7. Drinking Water Inspectorate. 2017. Drinking Water 2017 - Summary of the Chief Inspector's report for drinking water in England. http://www.dwi.gov.uk/about/annual-report/2017/Summary_CIR_2017_England.pdf. (2017). Accessed: 2019-03-28.
8. Zachary Parker, Scott Poe, and Susan V Vrbsky. 2013. Comparing nosql mongodb to an sql db. In *Proceedings of the 51st ACM Southeast Conference*. ACM, 5.
9. Roy F Spalding and Mary E Exner. 1993. Occurrence of nitrate in groundwater–a review. *Journal of environmental quality* 22, 3 (1993), 392–402.
10. Susanna TY Tong and Wenli Chen. 2002. Modeling the relationship between land use and surface water quality. *Journal of environmental management* 66, 4 (2002), 377–393.
11. Peter R Winters. 1960. Forecasting sales by exponentially weighted moving averages. *Management science* 6, 3 (1960), 324–342.