Project: Data Science
A study to assess the impact of nitrate concentrations on river water in Northern Ireland

Members
Adina Asim, Aaryan Kaushik, Mohammad Sazegar, Rahmanda Wibowo

Background
In the last two decades, there have been many reports in the British Isles of increased concentrations of nitrate in surface waters. In 1991, the EU introduced the Nitrates Directive, which aimed to reduce water pollution caused or induced by nitrate from agricultural sources [1]. This is important because high nitrate levels in water (> 11.259 mg NO3-N/L) can cause serious health effects on human and aquatic ecosystems. For example, excessive concentrations of nitrate in water can cause a rapid growth of algae population (algae bloom) that blocks sunlight through the water, reduces oxygen levels and produces toxins in the water. And also because in many places, surface waters are used as the main source of drinking water, uncontrolled nitrate levels in these parts can have a dangerous effect for a large number of people.

There have been many studies in the past that observed a high level of nitrate in water that was caused by high precipitation of rainfall. After some rainfall events, soil pores can transport water and carry some nitrate concentrations to the river [2]. Due to the fact that 75% of lands in Northern Ireland is used for agriculture, it can be an interesting study to analyze if there is a relationship between high precipitation of rainfall and the level of nitrate in river water in this specific region.

Objective
This study aims to analyze the association between rainfall and nitrate concentrations (NO3-N) in river waters, and to assess its overall impact on the quality of river water in Northern Ireland.

Challenges
At the start of the project, we decided to choose the topic on finding the association between rainfall and nitrate concentrations. Based on some consultations, we tried to find the dataset from the agricultural region of Montana and Mississippi in the US. However, the dataset from these regions were not usable due to the high number of missing data and very limited number of observations. After spending a lot of time finding nitrate dataset from the US without a good result, we decided to expand the search outside the US. Eventually, we found the nitrate dataset from India and Northern Ireland, and we decided to use Northern Ireland's dataset because the dataset was better in terms of completeness and source reliability.

We faced a different challenge when searching for rainfall dataset. Logically, rainfall datasets should be more accessible because weather forecasting has been well developed since a long time ago. However, public rainfall datasets were actually not that common on the internet. Most of the public datasets were coming from satellite and in the form of gridded format which requires a significant time and effort to pre-processing them into usable datasets for analysis. We thought that it was not feasible to do very complex data

management within a very short time. After spending more time, we found a free historical weather API from open-meteo.com. There was a limitation on this historical weather API. It used 30 kilometre resolution of the gridded data and therefore cannot represent very small scale patterns in precipitation. Thus, there might be significant differences to local measurements. Regardless, we decided to use this API and accept the limitation because it was easier to work with so that we could continue the study.

More than 65% of the time budget of this project was spent on dataset discoveries and data management, and thus the analysis that can be done was limited. However, we learned many lessons and obtained new insights from doing this 2 weeks hackathon.

Dataset

The water quality dataset was retrieved from DAERA, UK [3]. The data was repeated nitrate measurements from various monitoring sites in Northern Ireland from 1990 to 2018, covering 141,431 rows and 17 variables in a JSON format which was converted into a CSV file. At the start of the study, the data was only recorded from 129 monitoring sites. From the year 2000 onward, 624 sites were present, but most of them had missing NO3-N (5% missing values of NO3-N). We wanted to see how many sites had observations every month over 19 years, and we found out only 94 sites had completed the study between 2000 and 2018 (21354 observations).

Rainfall dataset was retrieved from open-meteo.com historical weather API which is based on ERA5 grid dataset [4]. To enrich the water quality dataset with this rainfall data, first we summarised the longitude, latitude, and the first and last date measurement of NO3-N for each monitoring site. Then, we run the script to make multiple batches of API requests with these parameters to open-meteo.com, export them as multiple CSV files, and perform some data manipulation techniques to combine these rainfall data with the water quality dataset. The data description of the final data can be found in Table 1.

Table 1. Data description of the final dataset.

Variable	**Description**
Site Code	Code for river monitoring sites (character)
latitude & longitude	Coordinates of river monitoring sites (numeric)
Date	Date of the measurement (date string with format yyyy-mm-dd)
NO3_N_MGL	Nitrate nitrogen level (mg/L) in the monitoring site (numeric)
rainfall	Amount of rain (mm) at the day of the measurement (numeric)

Methodology

Initially, we identify in which regions very high or very low values of nitrate and rainfall lie. This assessment was made with the help of hotspots (regions where the concentration is high) and coldspots (regions where the concentration is low). To observe the cluster pattern

of nitrate concentrations and rainfall in the given regions, K-means clustering was performed. Finally, to draw a marginal inferences by taking into account multiple measurements for each site given in the study, a linear mixed model has been taken into account. We estimated random and fixed effects for mean nitrate in a month for each site specifically. The statistical model is formulated as follow:

$$log(Y_{ij}) = \beta_0 + \beta_1 MeanRain_{ij} + b_{0i} + b_{1i} MeanRain_{ij} + \epsilon_{ij}$$

where $i = 1, ..., n$ is the monitoring site index, $j = 1, ..., 12$ is the month, Y is the NO3-N outcome (log transformed), β_0 and β_1 are the fixed effect for their respective regressors, $b0$ and $b1$ are random intercept and random slope for mean rain, and $\epsilon_i \sim \mathrm{N}(0, \sum_i)$.

The log transformation is needed to reduce the skewness of the outcome distribution.

We argued that time has any effect on nitrate level in the river water because of the dynamic nature of the river water. The time information in the dataset is mainly used to link between the nitrate level and the weather condition at that time, and hence we do not consider time for inferences.

Software

We used multiple softwares to support this project. We used R version 4.1.0 for data wrangling and data analysis, Python version 3.9.12 for data fetching and wrangling, Tableau for spatial visualization and SAS version 9.4 for statistical modelling.

Data Visualization

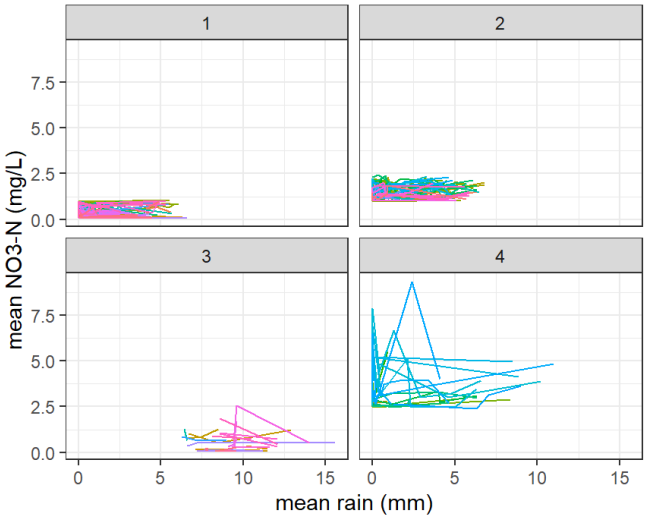
Figure 1 in the poster illustrates the daily average of NO3 (mg/L) and amount of rainfall (mm) for each month of the year among all stations between the years 2000 and 2018. As it shows there is a positive relation between the amount of rainfall and NO3 for the locations with moderate rainfall, however, locations with really high amounts of rainfall do not have high amounts of NO3 respectively. It seems that other factors such as topography are affecting this relationship for specific locations. Figure 2 in the poster is illustrating the topography of the country, most of these locations are located on low levels of the ground next to the mountains. Considering additional information such as topography (level of the ground) and agricultural cycles could result in more accurate representation of the relation between rainfall and NO3.

Figure 3 in the poster shows the fluctuation of NO3-N in 2018 in each site. It is observed that there is a highly increasing trend to the maximum level of NO3-N at the end of the year, and this coincides with the fact that around October and January is the wettest period in Northern Ireland. Figure 4 on the left side in the poster shows the result of k-means clustering in 2018 observations with 4 clusters. From the plot, we can see that cluster 3 and 4 are interesting. Cluster 3 has the highest mean of rain of all clusters, but the mean range of NO3-N is the lowest one. Cluster 4 has a wide mean range of rain, but the mean NO3-N is very high. Figure 4 on the right side in the poster displays the clusters based on geographical location. Cluster 4 observations are gathered in the bottom right of Northern Ireland, where other clusters seem to spread out in the map.

Linear Mixed Model

Individual site plot between dependent and independent variables shows the variability of the data that can be used as the validation basis of our mixed model formulation. It can be observed from Figure 5 that there is a high variability between sites at the intercept and slope, even though there is no clear linear relationship between the rainfall and the NO3-N concentration. We performed a likelihood ratio test to compare three models: with no random effect, with only random intercept, and with random intercept and random slopes. The result indicates that inclusion of random intercepts ($\chi 2 = 23125.1$, $p = 0.0015654023$) and random slopes ($\chi 2 = 14163.1 =$, $p = 0.006737947$) are significant at 5% of significance level. Hence interpretations are based on the full model.

Figure 5. Individual site plot mean rain vs mean NO3-N in 2018 per cluster.



After fitting the final linear mixed model, we can interpret the fixed effect to draw the marginal conclusion which was given in Table 2. Both intercept and mean rain effect was significant at 5% level of significance. This means that for every increase of 1 mm of average daily rain in a month causes an increase of exp(-0.01105) = 1 mg/L of nitrate level in river water.

Table 2. Result of linear mixed model.

Effect	**Estimate**	**Std. Err**	**Pr > |t|**
Intercept	-0.3937	0.1132	0.0008
mean_rain	-0.01105	0.001667	<0.0001

Conclusion & Discussion

- From the hotspots analysis, it is observed that the region with a moderate level of rainfall also has a moderate level of nitrate in water. The region with a high level of rainfall has a low level of nitrate, but around these specific regions the nitrate level is high in a distance. Interestingly, cluster 4 from k-means consists of these regions.
- Without including coordinates information in k-means, the clusters still show the topography differences of Northern Ireland.

- Based on the linear mixed model result, it is shown that there is a positive relationship between the average daily rain and the nitrate level in river water.
- Spatial analysis method is another option to analyze this type of dataset to take into account the correlation between clusters of regions.
- Cluster 3 and cluster 4 have different characteristics in terms of nitrate level and rainfall level. Taking into account these differences by adding topography information to the dataset might provide more meaningful insight to the study.

References

1. Northern Ireland Statutory Rules (2019). The Nutrient Action Programme Regulations (Northern Ireland) 2019 No. 81. https://www.legislation.gov.uk/nisr/2019/81/contents/made.
2. Jabloun, Mohamed & Schelde, Kirsten & Tao, Fulu & Olesen, Jørgen. (2015). Effect of temperature and precipitation on nitrate leaching from organic cereal cropping systems in Denmark. European Journal of Agronomy. 62. 55–64. 10.1016/j.eja.2014.09.007.
3. DAERA UK (2022). River Water Quality Monitoring 1990 to 2018 - Nitrate, Northern Ireland. Accessed 15 December 2022. https://opendata-daerani.hub.arcgis.com/.
4. Hersbach et al (2018). ERA5 hourly data on single levels from 1959 to present. Copernicus Climate Change Service (C3S) Climate Data Store (CDS). (updated daily), 10.24381/cds.adbb2d47.