Prediction of Ammonium concentration in a river using unsupervised Machine Learning

Vincenzo Guzzardi [1]

Abstract

Pollution in rivers are everywhere and it needs to be monitored. Unsupervised Machine learning is useful here to characterize and predict the concentration of the pollutant. Here NH_4 monitoring stations are used to create a concentration regime in the river. To do that, Kmeans is used over time and space. The results show that most of the river is at low regime. Higher concentrations can be seen downstream and after 2011.

1. Introduction

Ammonium nitrogen (NH_4) is well known as a pollutant from waste disposal and contaminated area (3). It's a toxic compound that is important to monitor in rivers water. The problem is mainly the eutrophication of the river water and is harmful for fishes and other species (2). Monitoring stations are often use to measure the concentration at a specific time and space. It allows to have a real value, but stations can be broken and it's not always easy to have a representative measurements. So, it would be interesting to predict concentration over space and time. Machine Learning algorithm are a good tools to achieve this goal because it's easy and cheap to set up.

So, the aim of the project is to determine a regime of concentrations of ammonium nitrogen in a river by using monitoring station to find a solution to know where and when the concentrations are the most problematic. So, we can ask the following question: what are the regime of ammonium concentration in the time and space of the river?

Table 1. Name of attribute of dataset

ATTRIBUTE NAME	DESCRIPTION
ID STATION	IDENTIFICATION NUMBER OF STATION
DATE	DATE OF MEASURE [DD.MM.YYYY]
NH_4	CONCENTRATION OF AMMONIUM MEASURED [MG/L]
DISTANCE	DISTANCE FROM THE SOURCE [KM]

2. Methods

2.1. Data

The dataset from the southern Bug river in Ukraine with 21 monitoring stations was used.

The attribute presented on the table 1 shows the information available for this set of data. All were used for this project.

The data are monthly or quarterly measured and can vary between the station. It covers 800 km of river during 26 years since 1993. There is 3499 measures of concentration with duplicates. After removing them there is 3436 measurements that will be use for the analysis.

The data can be found at this link: https://www.kaggle.com/datasets/vbmokin/ammonium-prediction-in-river-water?resource=download

2.2. Preprocessing

At first, the CSV file is imported as a Pandas DataFrame to work with. Then, the date is converted into monthly time series. To do it, the dates are changed into a Datetime format and change into months and duplicates between station and dates are removed. Now, the goal is to reshape the data to have a matrix with the months on the rows and the stations on the columns. The stations also give the distance. So, we have the information about time and space depend if we take the rows or the columns. To feed the Kmeans, all the cells have to be fill with a value. But some stations begin the measurements later than the first date or are not consistent.

[1] Master of environmental sciences, University of Lausanne, Lausanne. Correspondence to: Vincenzo Guzzardi <vincenzo.guzzardi@unil.ch>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021. Copyright 2021 by the author(s).

It results with a lot of NaN values. There were filled with the median of the corresponding station.

2.3. Unsupervised clustering

First, the concentrations will be considered on the time scale. So, the number of sample will be the month and the number of features will be the stations. Kmeans form scikitlearn is used to create 3 regime of the river. After the same will be done to see the concentration in the space instead of time. To do that the matrix is transposed to inverse the samples and features. Different numbers of cluster (from 2 to 6) are tested to see which model represent the best the concentration regime.

2.4. Metrics

The silhouette and inertia metrics are used to evaluate the model and test the best cluster number. It was done only for the time scale because the metrics were not able to perform a high number of features (322 months). The inertia can show the best k if there is an elbow but it's not the best ways to find k number. The silhouette score balances the distance of the centroid with the cluster. More there is clusters, more the points will be near a centroid. This metric take this in account. So it is a better ways to have the best number of cluster.

As it is an unsupervised clustering, there is no table of accuracy. The reason is that there is no label to test the accuracy of the cluster by this way. For the same reason, the data aren't split into training and test set with this kind of method.

The code is available at this link : https://github.com/VGuzz/2022_ML_Earth_Env_Sci/blob/main/Project_ML2022.py

3. Results

To visualize the distribution of the stations along the river, the Figure 1 shows the stations with the distance from the source. The stations are relatively well distributed along the river.

The Figure 2 presents a time serie of three stations to have an idea of the distribution of the concentrations. The data before and after the fill of the NaN value show that there is a lot of NaN. The values are very similar for some stations after putting the median value.

The clustering with the Kmeans algorithm was first do with $k = 3$ to see the representation of the regime. The result can be seen in the Figure 3 for the time and the Figure 4 for the space.

It is interesting to see that higher concentration are mostly after 2011 and at the end of the river. This observation is the same as the time series of Figure 2. Indeed, there is more measurements since around 2006, so a better representation of the data. However most of the river are at a low concentration with a regime of 0. It can depend on many things detailed in section of discussion.

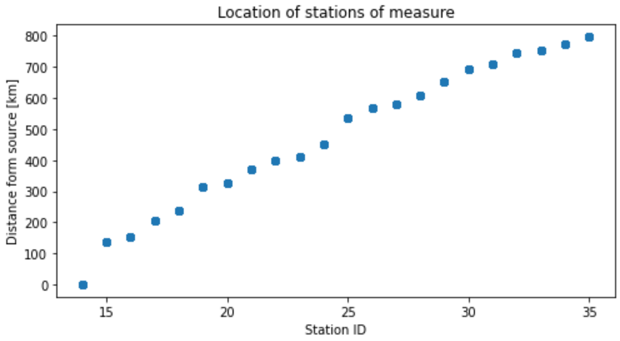

Figure 1. Locations of monitoring station

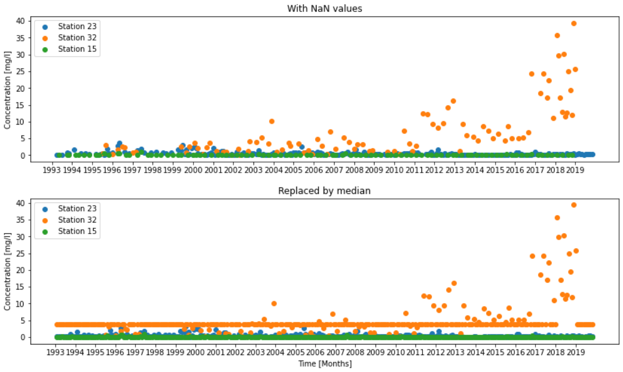

Figure 2. Concentration of three stations over time before and after filling NaN values

Concerning the metrics, they are presented in Figure 5. The silhouette metric show that the best number of cluster is 2. The elbow of inertia is located at 3 cluster. It is not the same number but it is relatively close. There is not better number of k for the silhouette and the elbow is well marked.

Globally, the results show that most of the river is at a low regime of concentration of ammonium with some pics downstream and after 2011.

4. Discussion

There are some interesting results to analyse. There is certainly one or several sources of NH_4 after around 700 km at station 31. The river goes through some city area, so

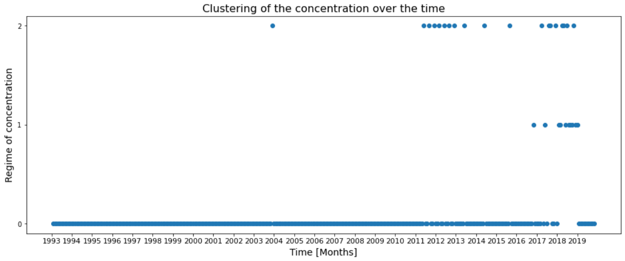

Figure 3. Regime of the river over time

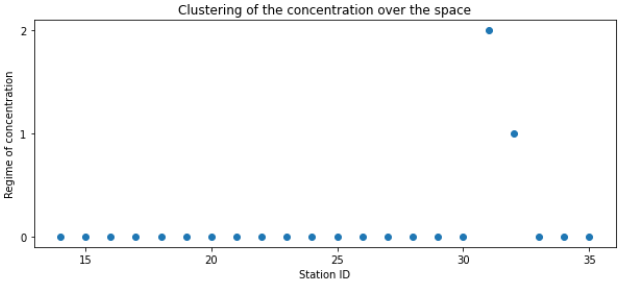

Figure 4. Regime of the river over space

it can explain the increase of the concentration (1). Bacteriological activities can change the concentration in the river, by degrading the ammonium too (4). Here there is no marked seasonality on the concentration of the ammonium in the river water. A more detailed analysis of the context could improve the interpretation of the results. Especially the reason of the location and the date of the beginning of the higher concentrations.

The metrics are interesting, because the best cluster model is with only 2 clusters. It is meaningful because most of the data set is at a low regime and only few concentration are beyond the majority of concentration. The number of cluster from 3 to 5 have the same score, but the k=6 goes down. So it don't explain the data at all. The Figure 3 and 4 show however 3 clusters to see more details despite it doesn't explain much than 2 clusters.

Concerning the results of the unsupervised clustering, the results are not really informative about the reality. As seen with the metrics, the low regime shows that the river has lower concentration. In fact, there is a lot same values, so it will be classified as the same number of cluster logically. Despite the fact that there is 3436 measurements, after the reshape of the data there is almost the same number of NaN values. So, the clustering is not really helpful in this case.

The dataset has a lot of station that there is no measurements during a given time. For example, some stations began there monitoring in 2016 but the measurements began in 1993. With the matrix there is a lot of NaN values, so the clustering will be biased. Maybe, this kind of clustering is not a good solution if there is not a homogeneous data over time as seen in Figure 2.

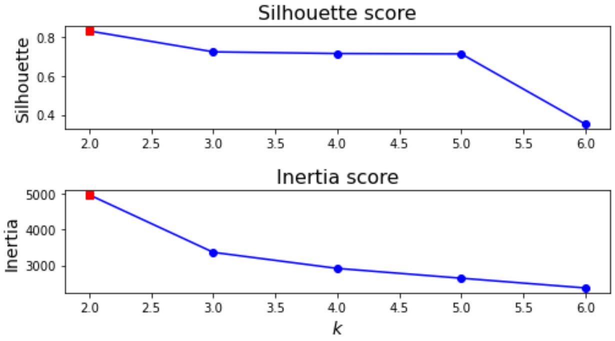

Figure 5. Silhouette and inertia metrics of Kmeans

5. Conclusion

At the end, this project allows to predict the regime of the pollution of a river with unsupervised machine learning in time and space. It is possible after the classification to see the concentration of each regime and compare it with threshold from environmental agencies for example. However it is really sensitive with the missing values. The metrics show that just 2 clusters explain the data but if there is more variability in the values, maybe the results would be more complex and interesting to see. If the measurements is continuous in time and space, some pattern could appear with this kind of project to evaluate a seasonality or a point of interest in city like a water treatment plant. Unsupervised machine learning didn't work very well for this data but it could be good for another type of monitoring.

To conclude, it was interesting to see how the preprocessing is important to have a good results but also how to choose the good methodology. Indeed, it would be interesting for further research to test other type of methods to evaluate the feasibility of different machine learning algorithms on this dataset.

References

[1] Daoliang Li, Xianbao Xu, Zhen Li, Tan Wang, and Cong Wang. Detection methods of ammonia nitrogen in water: A review. *TrAC Trends in Analytical Chemistry*, 127:115890, June 2020.

[2] Daniel C. V. R. Silva, Lucas G. Queiroz, Rodrigo J. Marassi, Cristiano V. M. Araújo, Thiago Bazzan, Sheila

Cardoso-Silva, Gilmar C. Silva, M. Müller, Flávio T. Silva, Cassiana C. Montagner, Teresa C. B. Paiva, and Marcelo L. M. Pompêo. Predicting zebrafish spatial avoidance triggered by discharges of dairy wastewater: An experimental approach based on self-purification in a model river. *Environmental Pollution*, 266:115325, November 2020.

[3] Wangshou Zhang, Dennis P. Swaney, Bongghi Hong, Robert W. Howarth, and Xuyong Li. Influence of rapid rural-urban population migration on riverine nitrogen pollution: perspective from ammonia-nitrogen. *Environmental Science and Pollution Research*, 24(35):27201–27214, December 2017.

[4] Mei-ai Zhao, Hao Gu, Chuan-Jie Zhang, In-Hong Jeong, Jeong-Han Kim, and Yong-Zhe Zhu. Metabolism of insecticide diazinon by Cunninghamella elegans ATCC36112. *RSC Advances*, 10(33):19659–19668, May 2020. Publisher: The Royal Society of Chemistry.