MANIPAL INSTITUTE OF TECHNOLOGY
MANIPAL
A Constituent Institution of Manipal University

Cluster Analysis of Cyanobacteria Growth

By:

Prajwal Prabhu : 210953002

Joshita Bolisetty : 210953070

Shubhanshu Verma : 210953072

Index

INTRODUCTION

Cyanobacteria, also known as blue-inexperienced algae, are a various group of photosynthetic bacteria that play a vital function in aquatic ecosystems. They are primary manufacturers, meaning they convert sunlight into electricity via photosynthesis, and they shape the base of the food chain for many aquatic organisms. However, cyanobacteria also can be harmful, causing blooms which could produce pollutants that could damage humans, animals, and aquatic lifestyles.

The distribution of cyanobacteria in aquatic ecosystems is stimulated by means of a range of things, which includes water temperature, nutrient availability, and light availability. Understanding the conditions which are conducive to cyanobacteria boom is vital for predicting and handling cyanobacteria blooms.

Clustering-based totally evaluation is a statistics mining method that may be used to pick out patterns in facts. This method may be used to pick out situations and areas which might be conducive to cyanobacteria increase.

In this report, we are developing a clustering-based evaluation framework for cyanobacteria distribution in aquatic ecosystems. We will use this framework to identify situations and areas which might be conducive to cyanobacteria growth.

Objectives

The objectives of this report are to:

- Develop a clustering-based analysis framework for cyanobacteria distribution in aquatic ecosystems.
- Use this framework to identify conditions and regions that are conducive to cyanobacteria growth.

LITERATURE SURVEY

Abdullah *et al.* [1](2022) The study addresses the critical issue of assessing and managing COVID-19 risk in Indonesian provinces during the early stages of the pandemic. The challenge is to determine the proximity or similarity between provinces based on the number of confirmed COVID-19 cases, recovered cases, and deaths, crucial for informed decision-making and policy formulation. The research method involved data collection from the Indonesian COVID-19 Acceleration Task Force website, categorization into 3three categories: confirmed cases, recovered cases, and deaths, and utilization of the K-Means Clustering method. Advantages include efficient identification of patterns and clear insights into provincial similarities, while disadvantages include sensitivity to initial cluster centroids and the need for prior knowledge of the desired number of clusters. The primary findings revealed three distinct clusters of provinces, providing valuable insights for policy formulation. In conclusion, the study's application of the K-Means Clustering method demonstrates its potential for data-driven decision-making in managing COVID-19, offering significant implications for government policies and future predictions based on provincial data.

Shutaywi, *et al.* [2](2021) The assessment of clustering methods in machine learning across diverse domains is pivotal yet challenging due to the need for suitable algorithms and parameter selection, often reliant on labelled data for evaluation. To mitigate this, the study uses an approach utilizing the Silhouette index, an unsupervised metric measuring clustering quality based on cohesion and separation. This method integrates multiple kernel functions within kernel k-means clustering, allowing data projection into higher-dimensional spaces for accommodating complex cluster shapes. Its advantages include broadened applicability without labelled data and reduced sensitivity to kernel selection, yet limitations exist in its ability to capture all facets of clustering quality, its effectiveness can vary based on dataset chosen, and this method assumes the availability of multiple kernel functions, which may not always be readily accessible. Rigorous assessments via Monte Carlo simulations on benchmark datasets demonstrate the effectiveness of the proposed weighted clustering approach, primarily evaluated by the Silhouette index. Overall, this method marks a significant advancement by addressing challenges in clustering evaluation, presenting a valuable tool for enhancing clustering outcomes across applications while suggesting avenues for future refinements and real-world applications.

Yuan, *et al.* [3](2019) This study examines four key K-value selection algorithms—Elbow Method, Gap Statistic, Silhouette Coefficient, and Canopy—in K-means clustering. Each method, while offering distinct advantages, faces limitations. The Elbow Method seeks a sum of squared errors (SSE) 'elbow point' but struggles without a clear bend. Gap Statistic provides reliable results but becomes computationally demanding for large datasets. Silhouette Coefficient gives a balanced evaluation of cluster quality but faces complexity in distance matrix calculations. Canopy excels in handling extensive datasets with robustness to faults and noise. While all methods suffice for small datasets, Canopy stands out for its efficiency with larger and complex data. In conclusion, the Canopy method proves advantageous in computational efficiency, fault tolerance, and noise immunity for larger datasets, warranting further exploration with real-world multidimensional data for potential enhancements.

Vincent Cohen-Addad *et al.* [4] (2021) The paper addresses the fundamental challenge of clustering large datasets efficiently by proposing a novel coreset framework. The paper

introduces a coreset framework. The coreset is a smaller subset preserving clustering properties of the original dataset. It is constructed by iteratively adding data points to the coreset until the coreset is sufficiently representative of the original dataset. The advantage of this framework is being both highly efficient and effective in producing clusters akin to traditional algorithms, even with significantly smaller coreset sizes. However, challenges arise in selecting the right subset of data points to use as coreset despite provided heuristics. Evaluations across diverse datasets validate the framework's ability to generate clusters akin to traditional methods, highlighting its potential significance in clustering large datasets and its broad applicability in machine learning.

Ran, *et al.* [5](2021) This research paper addresses the challenge of identifying urban hotspots by introducing a novel K-means clustering algorithm augmented with a noise algorithm to detect and exclude outliers. This approach aims to enhance hotspot identification, crucial for various stakeholders. The advantages include that it is simple and easy to implement, is computationally quite efficient, making it suitable for large datasets. Whereas its disadvantage is that it is sensitive to the choice of the number of clusters(K) and it is not able to identify overlapping hotspots. necessitate consideration. Evaluation using Beijing's taxi GPS data demonstrates the algorithm's superior accuracy in hotspot identification compared to traditional K-means algorithms. Overall, this research introduces a significant advancement in urban data analysis, providing an efficient method for identifying urban hotspots likely to find broad applications in diverse fields.

Amin Shahraki, *et al.* [6](2020) This paper delves into clustering within wireless sensor networks (WSNs), aiming to group sensor nodes to enhance energy efficiency, scalability, reliability, and other metrics crucial in WSN performance. Reviewing a spectrum of clustering techniques based on objectives such as energy efficiency, scalability, reliability, quality of service, and security, it also explores network properties like mobility and heterogeneity. The advantages of this paper are that it provides a comprehensive and up-to-date overview of clustering techniques in WSNs and provides statistics on the chosen metrics. Whereas the disadvantage is that the paper does not go into much detail about the design and implementation of different clustering techniques. Analysis reveals energy efficiency as the predominant clustering objective, followed by scalability and reliability, mainly targeting static and homogeneous sensor networks. Despite its limitation in tutorial depth, the paper stands as a valuable resource for researchers and practitioners engaged in designing or implementing clustering algorithms for WSNs, providing an informative overview of techniques, objectives, supported properties, and performance metrics.

K. P. Sinaga [7](2020) The paper addresses limitations in traditional k-means clustering algorithms by proposing an innovative approach utilizing a weighted distance metric. This new algorithm aims to enhance efficiency and robustness, crucially reducing computational demands by assigning greater weight to data points closer to the centroid, thus minimizing the need for distance calculations. Additionally, its robustness to outliers is improved by assigning them to clusters farther from the centroid. The advantages include that the proposed algorithm demonstrates heightened efficiency and robustness through evaluations on various datasets, showcasing effectiveness across diverse data characteristics. But there is still a downside to it that the method requires the user to enter the value of K and finding an optimal value can be difficult. Overall, this algorithm presents a promising advancement in k-means clustering, offering enhanced efficiency and robustness, particularly beneficial when dealing with datasets of varying attributes and complexities.

A. Chhabra, *et al.* [8](2021) The paper addresses the challenge of fairness in clustering algorithms, highlighting their potential for unjust segregation based on sensitive attributes, impacting individuals and groups negatively. It reviews methods categorizable into pre-processing (altering data before clustering) and post-processing (modifying results after clustering) for achieving fairness. Pre-processing methods offer easier implementation but may disrupt data integrity, while post-processing methods are more flexible but often require new algorithm designs. Despite ongoing research on fairness-aware clustering algorithms, some studies indicate their ability to enhance fairness. The paper underscores the lack of consensus on fairness evaluation metrics, discussing common metrics like balance, isolation, and equalized odds. It emphasizes the need for further research on fairness-aware algorithms and novel fairness evaluation metrics, recognizing the pivotal role of fairness in clustering and the need for continued exploration in this domain.

D. Huang, *et al.* [9](2020) The paper tackles the computational complexity of spectral and ensemble clustering algorithms by introducing U-SPEC and U-SENC, focusing on scalability and robustness improvements thus addressing the limitations of the ensemble clustering algorithm(a technique that combines the results of multiple clustering algorithms). U-SPEC optimizes spectral clustering via representative selection and faster approximation methods, significantly reducing computational demands. U-SENC then integrates multiple U-SPEC clusters, enhancing robustness while retaining efficiency. Despite their scalability and robustness advantages, U-SPEC and U-SENC require more memory due to storing the entire affinity matrix, marking a drawback. Evaluations across large-scale datasets highlight their superior efficiency and robustness compared to existing algorithms. For instance, U-SPEC demonstrated clustering of 10 million data points in around 10 minutes, contrasting with existing spectral algorithms requiring hours or days for such tasks. These novel algorithms show promise in various applications like image clustering, social network analysis, and customer segmentation, offering improved scalability and robustness in large-scale spectral and ensemble clustering scenarios.

Khumalo, et al. [10](2020) The paper delves into the underexplored realm of Cytochrome P450 monooxygenases (CYPs) in cyanobacteria, highlighting their crucial role in secondary metabolite biosynthesis. Analyzing 114 cyanobacterial genomes, the study comprehensively identifies CYPs and secondary metabolite biosynthetic gene clusters (BGCs), unveiling the diverse nature of CYPs in cyanobacteria. While providing valuable insights into CYP diversity and their association with secondary metabolites, the study is confined to a limited set of cyanobacterial species, potentially overlooking CYPs present in unexplored species. Findings reveal a broad spectrum of CYP families in cyanobacteria, with a small fraction associated with secondary metabolite biosynthesis. Notably, the study identifies CYPs involved in terpene, polyketide, and non-ribosomal peptide biosynthesis, hinting at their potential for producing valuable secondary metabolites like antibiotics and pharmaceuticals. This exploration lays the groundwork for further research and development of novel methods harnessing cyanobacterial CYPs for secondary metabolite production, marking a significant advancement in cyanobacterial research.

METHODOLOGY

The methodology for this project will be as follows:

1. Collect data on cyanobacteria distribution in aquatic ecosystems.
2. Preprocess the data.
3. Apply clustering-based analysis to the data.
4. Interpret the results of the clustering-based analysis.

Expected Outcomes

- Identify conditions that are conducive to cyanobacteria growth.
- Identify regions that are conducive to cyanobacteria growth.
- Develop a framework that can be used to predict cyanobacteria blooms.

Data Preprocessing

During data preprocessing, we encountered missing values denoted by "NA" (Not Available) in some data entries. To make certain information consistency and dispose of potential mistakes, we decided to remove those entries with missing values.

Subsequently, we converted all the closing information entries to the "float" statistics type. This conversion become necessary to make sure a consistency and compatibility with our preprocessing and data analysis tools.

These information preprocessing steps have been vital in preparing the records for analysis and ensuring the validity of our findings. By getting rid of lacking values and changing all information to the "float" records type, we ensured that our analysis turned into based totally on clean, consistent, and reliable data.

Clustering-Based Analysis Algorithms

To pick out patterns in the information and uncover the factors influencing cyanobacteria distribution, numerous clustering-based evaluation algorithms may be hired. These algorithms effectively group records points primarily based on their similarities, permitting us to become aware of wonderful clusters of cyanobacteria distribution related to unique environmental situations.

Algorithm Selection

The choice of clustering algorithm depends on the characteristics of the data, the studies goals, and the preferred degree of granularity inside the analysis. Some of the clustering algorithms for cyanobacteria distribution analysis that we've used include:

1. K-means:

Centroid-based: K-means is a centroid-based algorithm where it partitions the data into K clusters, with each cluster represented by its centroid. It minimizes the sum of squared distances between data points and their assigned cluster centroids.

Number of clusters (K): One of the challenges with K-means is that you need to specify the number of clusters (K) beforehand, which might not be known in advance.

Sensitivity to Initial Centroids: The algorithm's performance can be sensitive to the initial placement of centroids, and it may converge to a local minimum.

Scalability: K-means is generally efficient for small to medium-sized datasets but might not be suitable for large datasets due to its iterative nature.

Assumes Spherical Clusters: K-means assumes that clusters are spherical and equally sized, making it less suitable for elongated or irregularly shaped clusters.

2. BIRCH:

Tree-based Hierarchical Clustering: BIRCH is a hierarchical clustering algorithm that uses a tree-like structure to represent the data. It incrementally builds a Clustering Feature Tree (CF Tree) to summarize the data in a compact manner.

Balanced and Incremental: BIRCH is designed to be scalable and can handle large datasets incrementally. It processes data in a single pass, making it suitable for streaming data.

Noise Tolerance: BIRCH is less sensitive to noise and outliers compared to K-means, making it robust in scenarios where the data might have irregularities.

Automatic Determination of Clusters: BIRCH automatically determines the number of clusters and does not require the user to specify K in advance.

Memory Efficiency: BIRCH is memory-efficient as it summarizes the data in a compact CF Tree, which allows it to process large datasets without storing all data points in memory.

3. Agglomerative Hierarchical Clustering:

Hierarchy of Clusters: Agglomerative clustering builds a hierarchy of clusters by successively merging or agglomerating individual data points or clusters.

No Need for Pre-specifying K: Unlike K-means, agglomerative clustering does not require the user to specify the number of clusters in advance. The hierarchy can be cut at different levels to obtain different clustering.

Linkage Methods: Agglomerative clustering uses different linkage methods (e.g., complete, average, ward) to determine how to merge clusters based on the distance between them.

No Sensitivity to Initial Points: Agglomerative clustering is less sensitive to the initial configuration, as it iteratively merges clusters based on a predefined linkage criterion.

Visualization of Hierarchy: Agglomerative clustering provides a dendrogram, which visually represents the hierarchy of clusters and their relationships.

4. Gaussian Mixture Models (GMMs):

Probabilistic Model: GMMs model the data as a mixture of several Gaussian distributions. Each cluster is represented by a Gaussian distribution, and a data point has a probability of belonging to each cluster.

Soft Assignment: GMMs use soft assignment, meaning each data point is assigned a probability of belonging to each cluster. This allows for more flexibility in handling data points that may belong to multiple clusters.

Can Model Elliptical Clusters: GMMs can model clusters with different shapes and sizes, making them more flexible than K-means in capturing complex structures in the data.

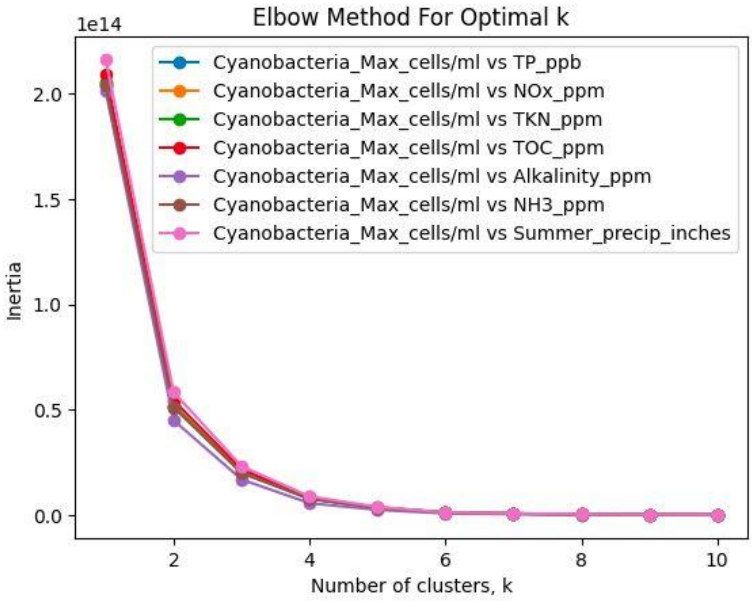
EM Algorithm: GMMs are typically trained using the Expectation-Maximization (EM) algorithm, which iteratively refines the parameters of the Gaussian distributions to maximize the likelihood of the data.

For algorithms where k is required to be predetermined, the **elbow method** is used. It is a heuristic used in clustering analysis to determine the optimal number of clusters in a data set. It works by plotting the within-cluster sum of squares (WCSS) for a range of cluster counts and selecting the point where the plot forms an "elbow" or a significant change in the rate of decrease. The WCSS is a measure of the total within-cluster variation, which is the sum of the squared distances between each data point and its cluster centroid.

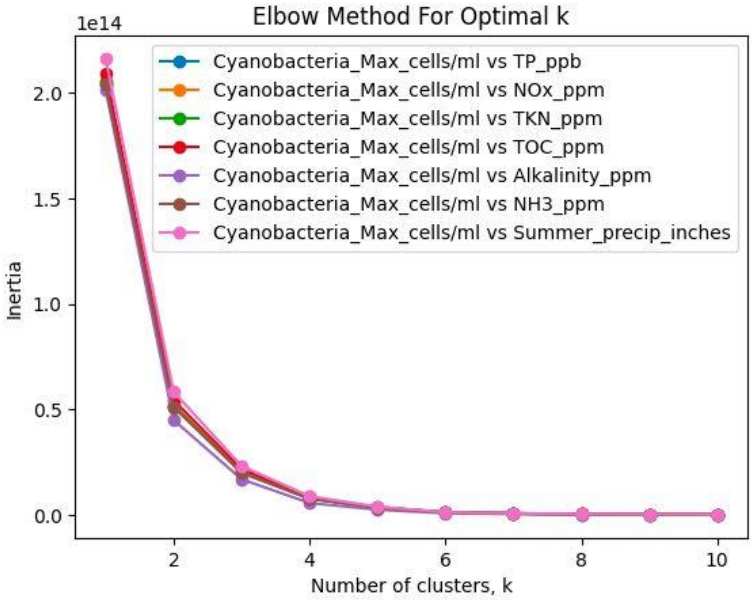
RESULTS AND DISCUSSION

Determining optimal K using elbow method

For dataset 1:



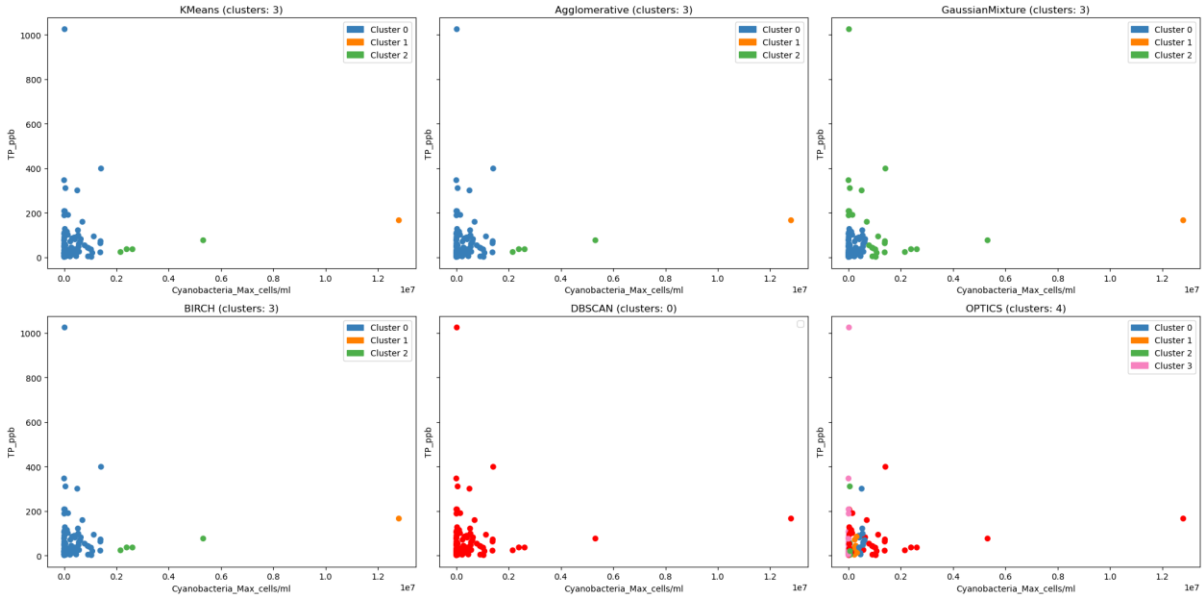
For dataset 2:



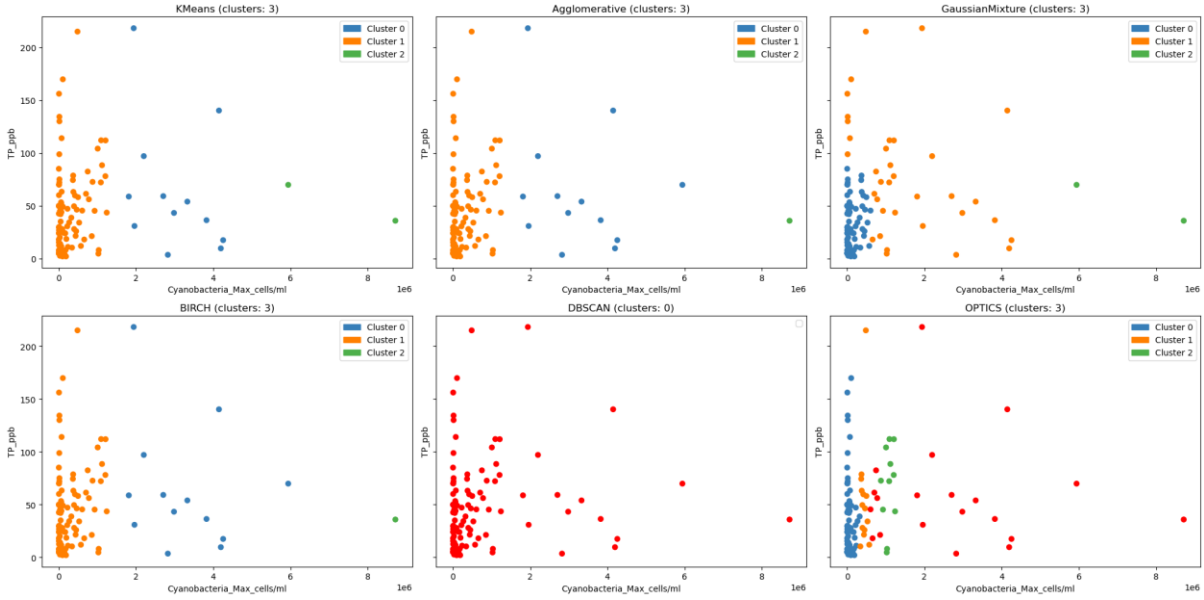
For both the datasets, the graph of the elbow method shows a clear bend at **k=3**. This suggests that the optimal number of clusters for both datasets is 3, meaning the data can be best represented by grouping it into 3 distinct clusters.

Cyanobacteria growth (cells/ml) vs Total Phosphorus (parts per billion)

For dataset 1:



For dataset 2:



The graph indicates a positive correlation among phosphorus and cyanobacteria growth. This signifies that as the quantity of phosphorus inside the water will increase, the quantity of cyanobacteria also will increase.

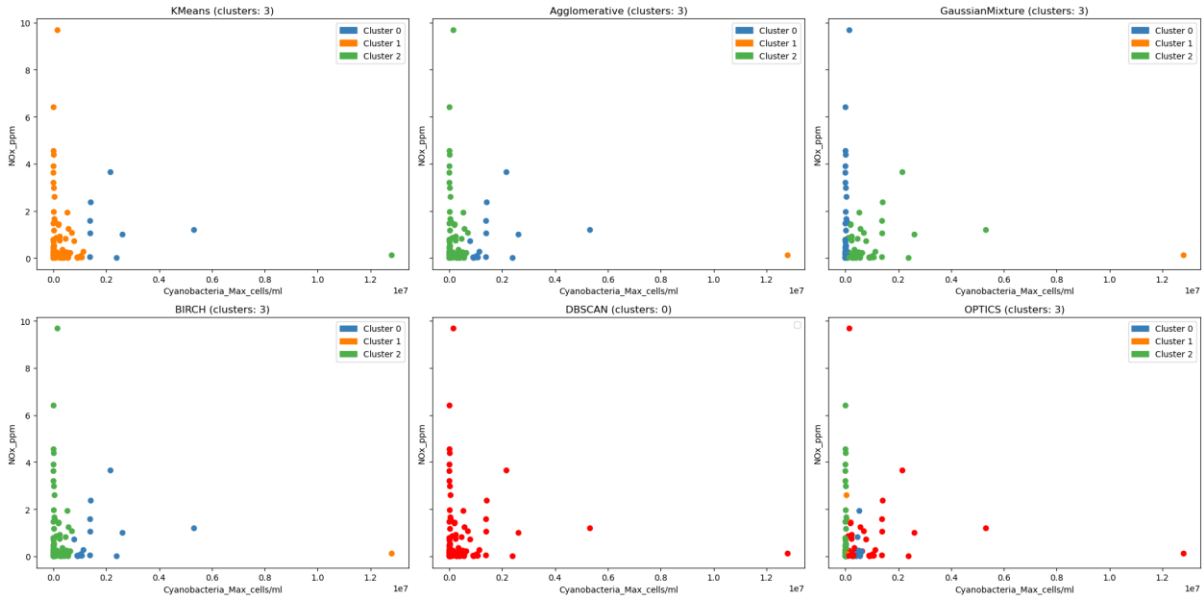
Phosphorus is a vital nutrient for cyanobacteria, as it is required for the synthesis of nucleic acids, phospholipids, and electricity-rich compounds along with ATP. When phosphorus is to be had in sufficient portions, cyanobacteria can develop and proliferate hastily. Studies have proven that growing phosphorus concentrations can cause enormous increases in cyanobacterial biomass.

This is why it is important to control the amount of phosphorus within the water with a view to prevent cyano blooms. Effective management of cyanobacterial blooms requires decreasing phosphorus inputs to aquatic ecosystems. This can be achieved through various techniques, which include:

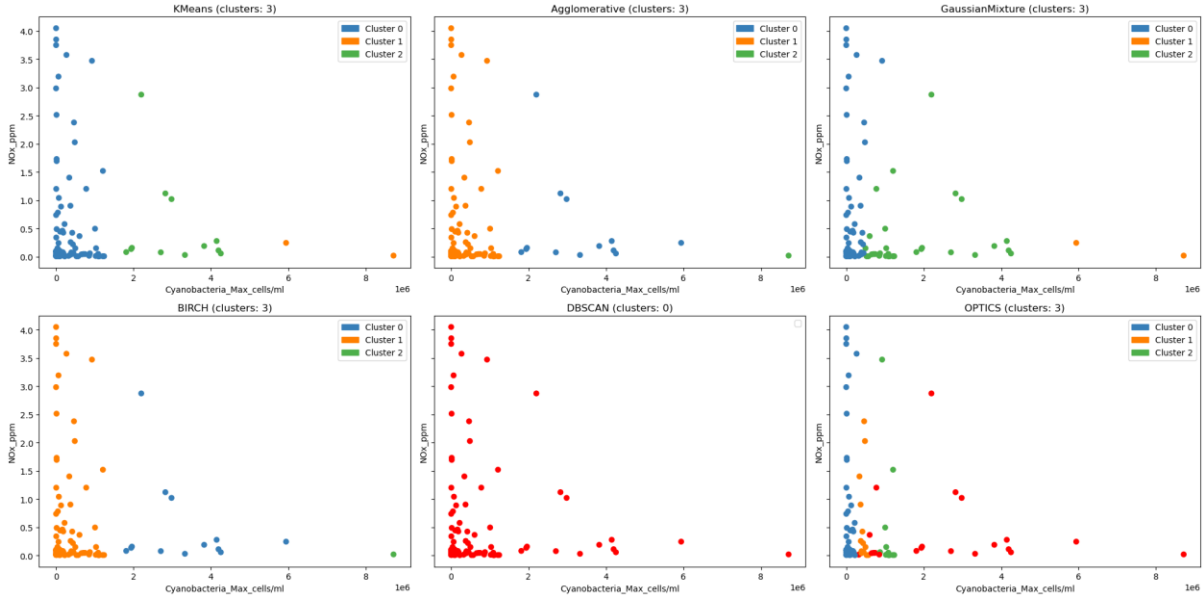
- Reducing agricultural runoff
- Improving wastewater remedy
- Controlling city runoff
- Reducing phosphorus in detergents

Cyanobacteria growth (cells/ml) vs Nitrogen Oxides (parts per million)

For dataset 1:



For dataset 2:



As seen from the graph, the relation between the concentration of Nitrogen Oxides and Cyanobacteria cells is inverse, that is whilst the attention of NO_x will increase, the Cyanobacteria cellular depend is discovered to be much less, whereas when the NO_x attention is low, the Cyanobacteria cells are discovered in excess.

Nitrogen oxide is a collection of gaseous compounds that consist of nitrogen monoxide (NO), nitrogen dioxide (NO_2), and nitrous oxide (N_2O). NO_x is fashioned thru herbal methods, consisting of lightning and volcanic eruptions, as well as human sports, including combustion of fossil fuels and industrial procedures.

Positive Effects: Nitrogen is a key aspect of amino acids, proteins, and nucleic acids, all of that are essential for cyanobacteria increase. When NO_x is present in sufficient quantities, cyanobacteria can develop and proliferate unexpectedly. Studies have proven that increasing nitrogen oxide concentrations can cause big will increase in cyanobacterial biomass.

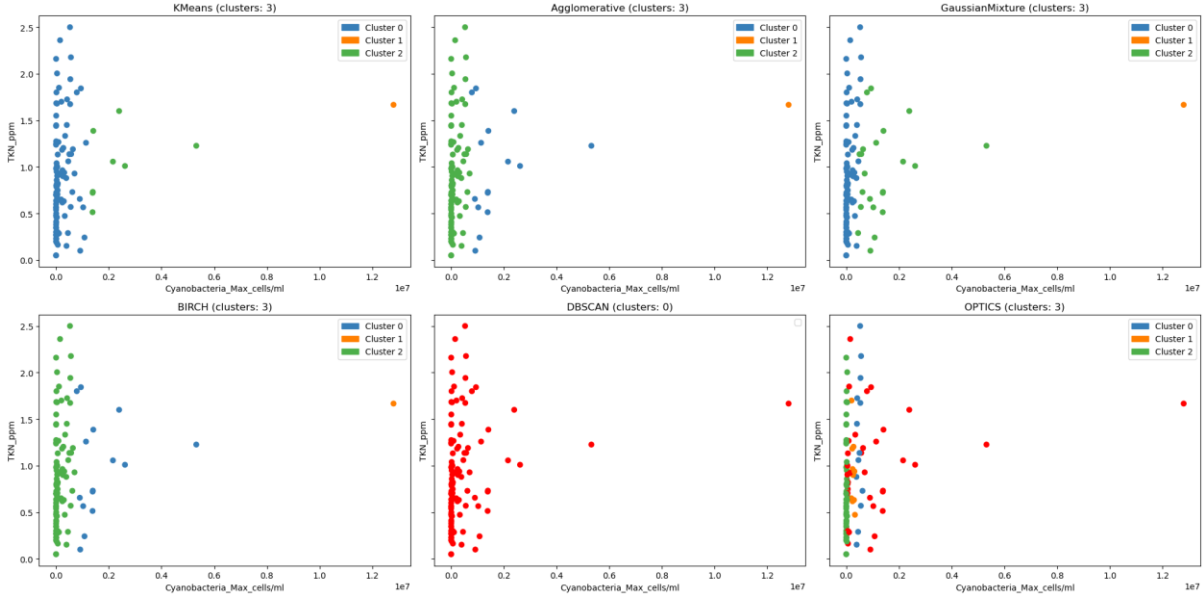
Negative Effect: Excessive nitrogen oxide degrees could have negative effects for cyanobacteria growth. High concentrations of NO_x can lead to the formation of reactive nitrogen species (RNS), which could harm cyanobacterial cells. Additionally, NO_x can make a contribution to acidification and eutrophication of aquatic ecosystems, which also can damage cyanobacteria populations.

The amount of Nitrogen Oxides can be kept optimal in water bodies by

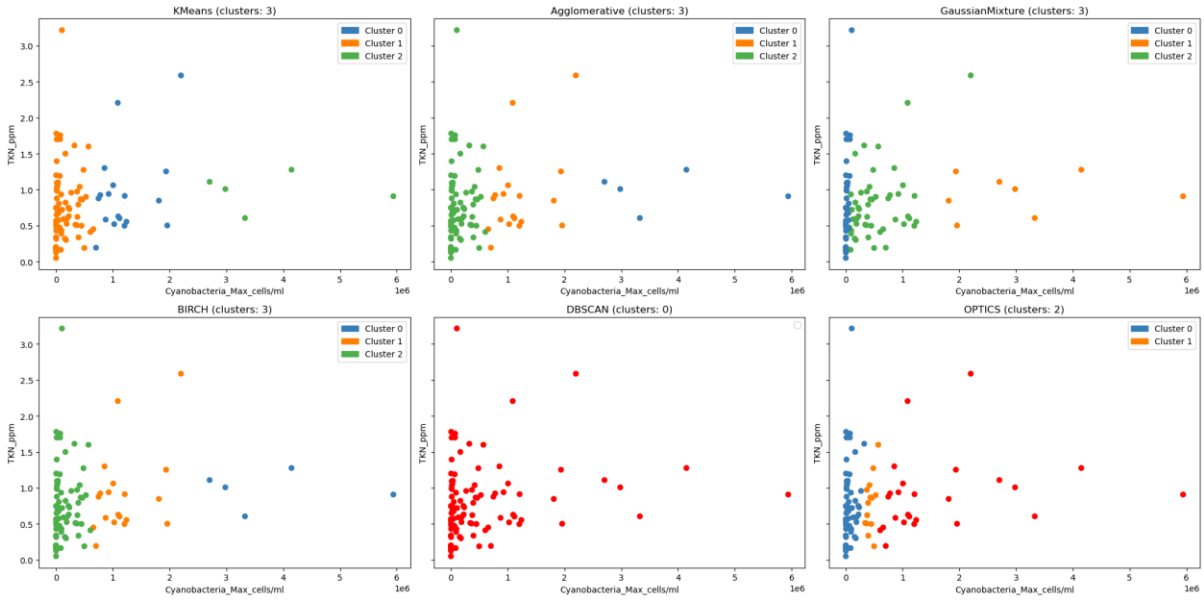
- Reducing Agricultural Runoff
- Improving Wastewater Treatment
- Controlling Urban Runoff
- Reducing Phosphorus in Detergents
- Restoring Riparian Buffers
- Managing Septic Systems

Cyanobacteria growth (cells/ml) vs Total Kjeldah's Nitrogen (parts per million)

For dataset 1:



For dataset 2:



As observed from the graphs obtained, as the Nitrogen concentration increases, the concentration of cyanobacteria cells also increase, hence they are directly proportional to each other.

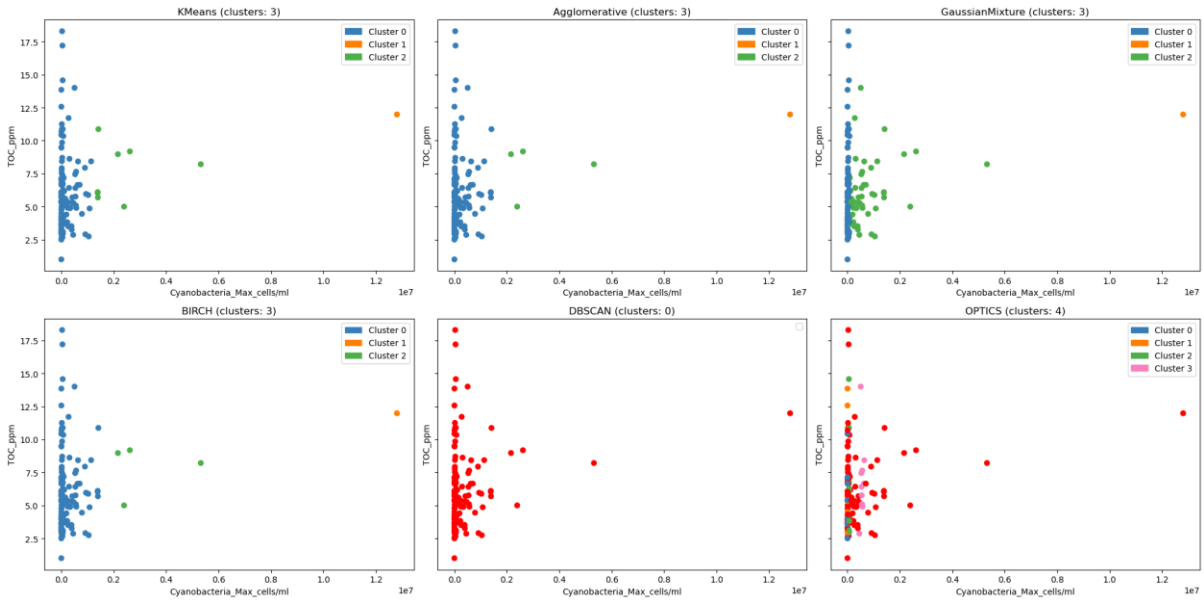
Nitrogen is an important nutrient for cyanobacteria, as it required for the synthesis of amino acids, proteins, and nucleic acids, all of which are vital for cyanobacteria increase. When nitrogen is in sufficient portions, cyanobacteria can grow and proliferate hastily. Studies have shown that growing nitrogen concentrations can lead to big increases in cyanobacterial biomass.

The availability of nitrogen can have a considerable effect on cyanobacteria boom. When nitrogen is constrained, cyanobacteria boom is slow, and their populations are small. However, when nitrogen is plentiful, cyanobacteria can grow swiftly, leading to blooms.

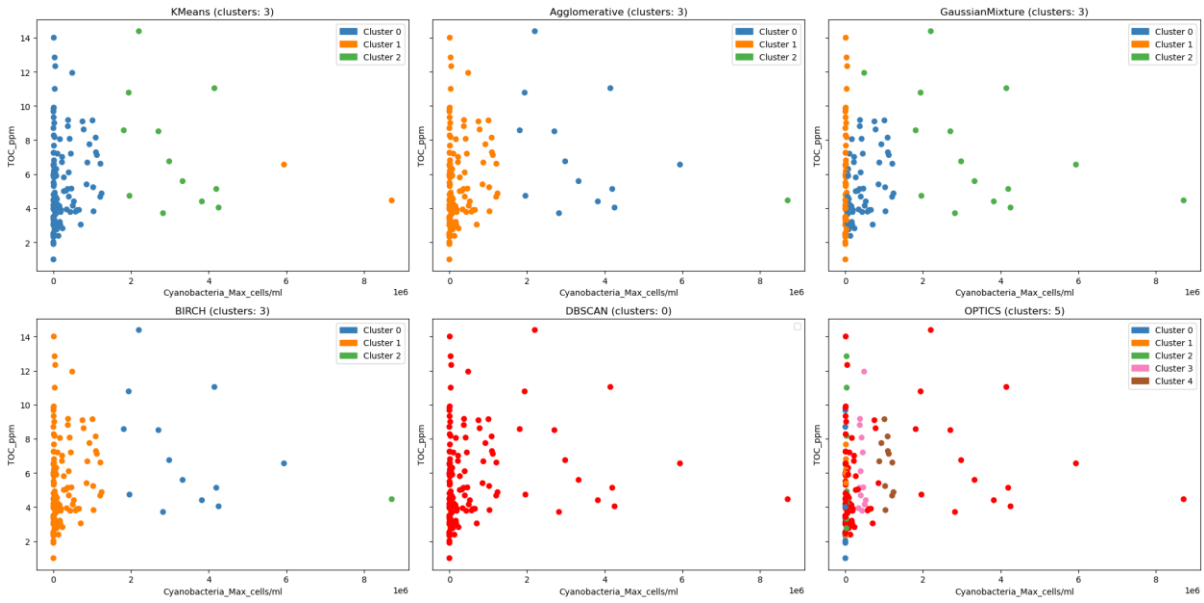
The methods of maintaining optimal amounts of nitrogen in lakes is similar to that of nitrogen oxides.

Cyanobacteria growth (cells/ml) vs Total Organic Carbon (parts per million)

For dataset 1:



For dataset 2:



The graph indicates that cyanobacteria can thrive in environments with high TOC levels. Additionally, at very high concentrations of carbon, the growth decreases.

Cyanobacteria can use TOC as a source of energy and nutrients. Cyanobacteria can produce EPS (Extracellular polymeric substances), which can bind to TOC and other organic matter. This can help to protect cyanobacteria from environmental stressors. TOC can also provide a habitat for bacteria and other organisms that cyanobacteria feed on. However, TOC can also stimulate the growth of other algae, which can compete with cyanobacteria for nutrients.

TOC can be controlled in water bodies by implementing the following:

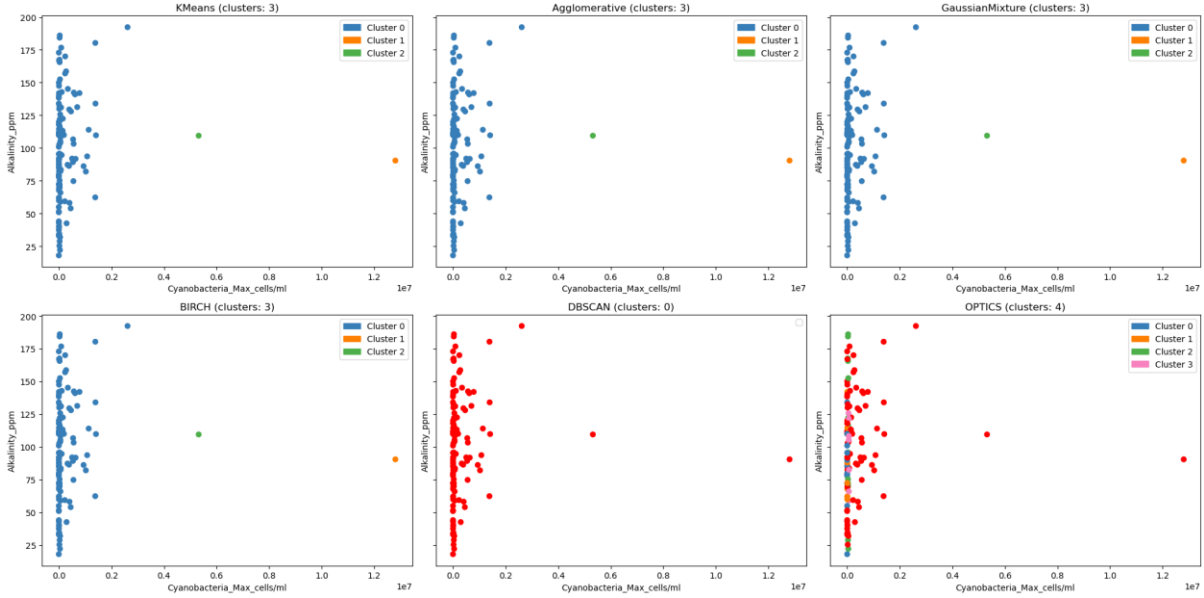
- Reducing runoff from terrestrial ecosystems
- Improving wastewater treatment
- Controlling nutrient pollution

For a safe environment, TOC levels in water bodies can also be on the higher side which vcan be achieved by:

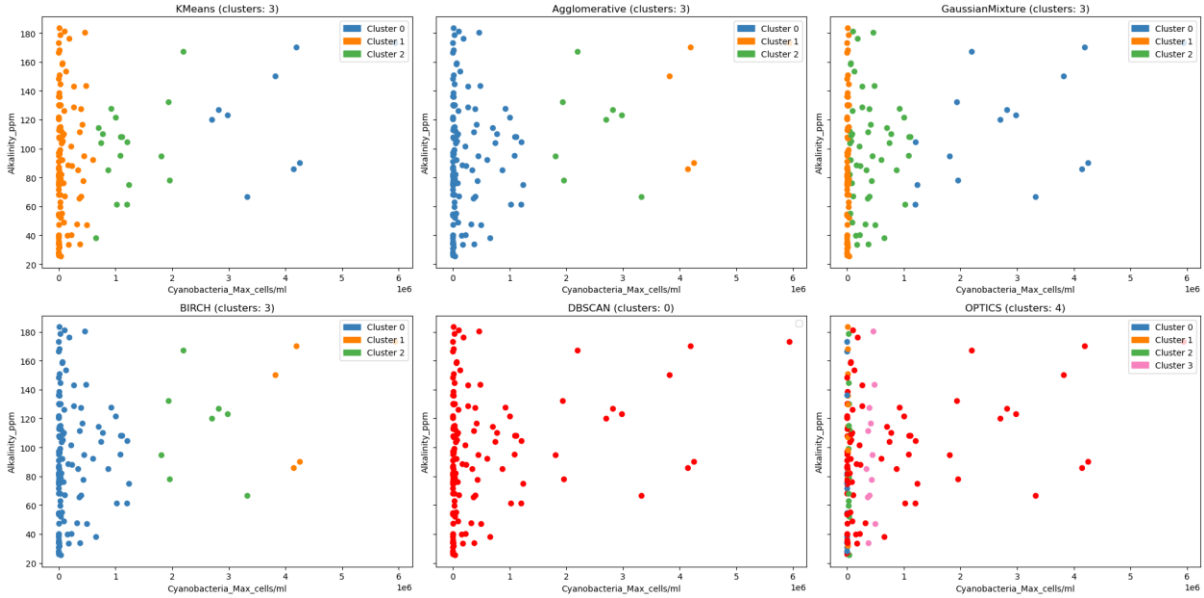
- Increasing vegetation
- Decreasing the removal of organic matter
- Adding organic matter
- Reducing sediment removal
- Increasing nutrient input
- Reintroduction of native species
- Minimizing human impact

Cyanobacteria growth (cells/ml) vs Alkalinity (parts per million)

For dataset 1:



For dataset 2:



As seen from the graph, the alkalinity of water needs to be an optimal range that is, neither too high nor too low to support maximum cyanobacteria growth.

Alkalinity is a measure of the capability to neutralize acids. It is expressed in terms of the concentration of bicarbonate (HCO_3^-) and carbonate (CO_3^{2-}) ions in the water. Alkalinity plays

a vital role in regulating the pH of aquatic ecosystems, and it can considerably influence the increase and proliferation of cyanobacteria.

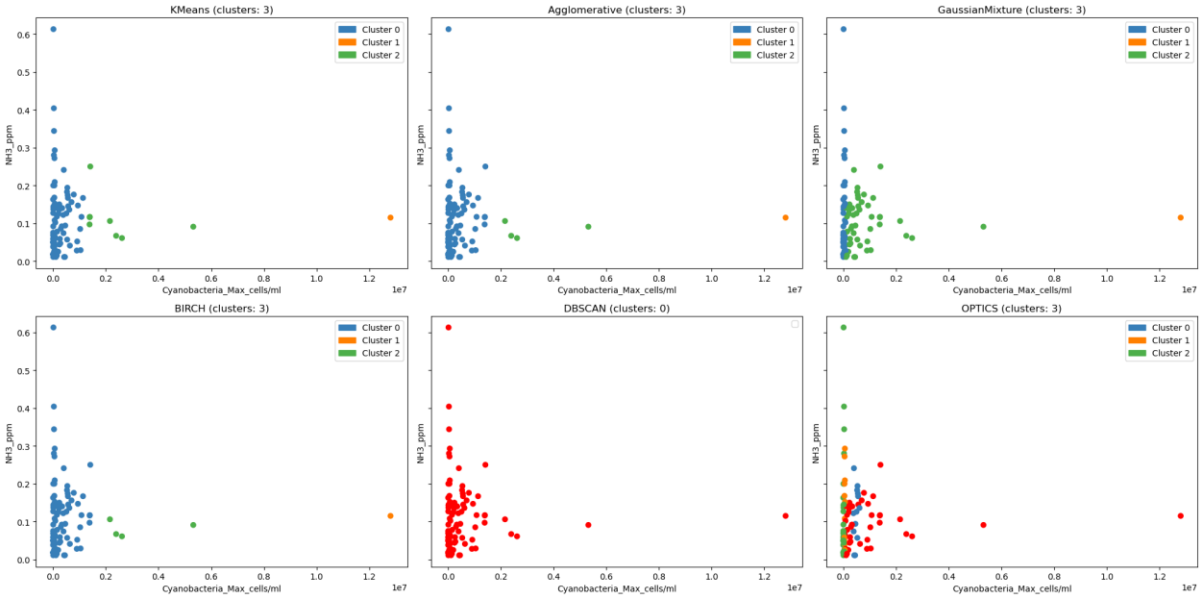
Positive Effects of Alkalinity on Cyanobacteria Growth: Cyanobacteria growth is supported in pH levels between 8 and 10. High alkalinity gives cyanobacteria numerous benefits, which include greater carbon fixation, suppression of toxic metals and so on.

Negative Effects of Alkalinity on Cyanobacteria Growth: While alkalinity typically favours cyanobacteria growth, excessively excessive alkalinity ranges can also have terrible outcomes like altered nutrient availability of a few nutrients, inducing physiological strain in cyanobacteria, susceptance to eutrophication.Managing factors that indirectly influence alkalinity can help control cyanobacteria growth:

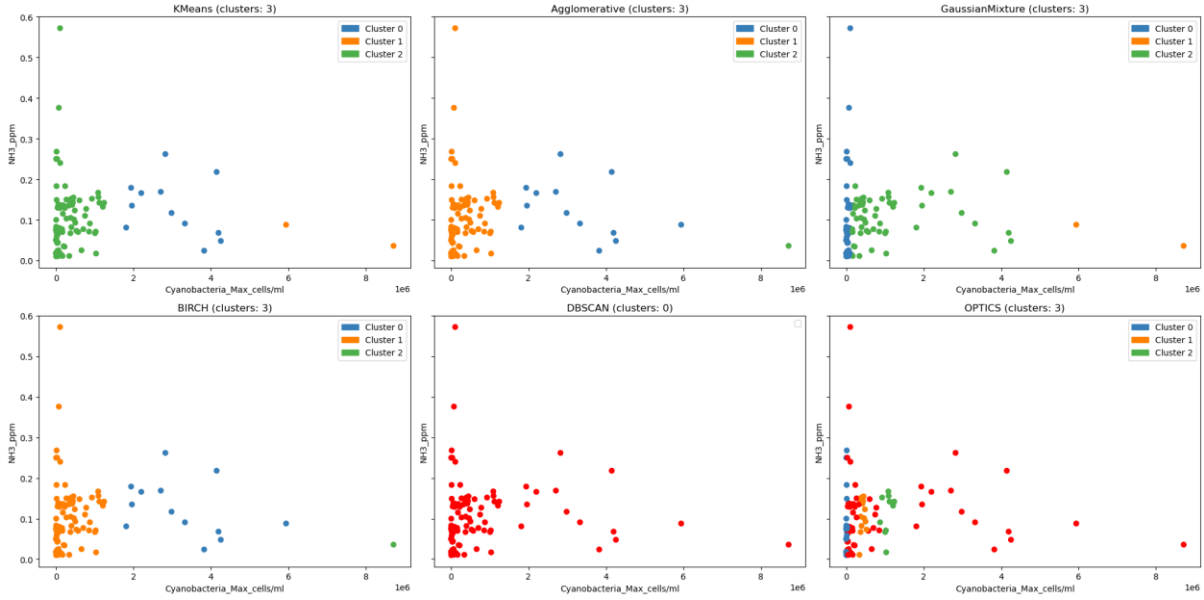
- Reducing nutrient inputs
- Controlling pH fluctuations
- Preserving natural buffering capacity
- Enhancing biodiversity

Cyanobacteria growth (cells/ml) vs Ammonia content (parts per million)

For dataset 1:



For dataset 2:



As seen from the graphs obtained, the maximum cyanobacteria concentration are found where the NH3 concentration is on the lower side. When the NH3 concentration is at the high, the Cyanobacteria concentration is significantly less.

Ammonia (NH3) is a form of nitrogen that is commonly found in aquatic environments. It is a primary nutrient source for cyanobacteria, playing a crucial role in their growth and metabolic processes.

Positive Effects of Ammonia on Cyanobacteria Growth include assimilation and increased cyanobacterial biomass production and faster growth rates.

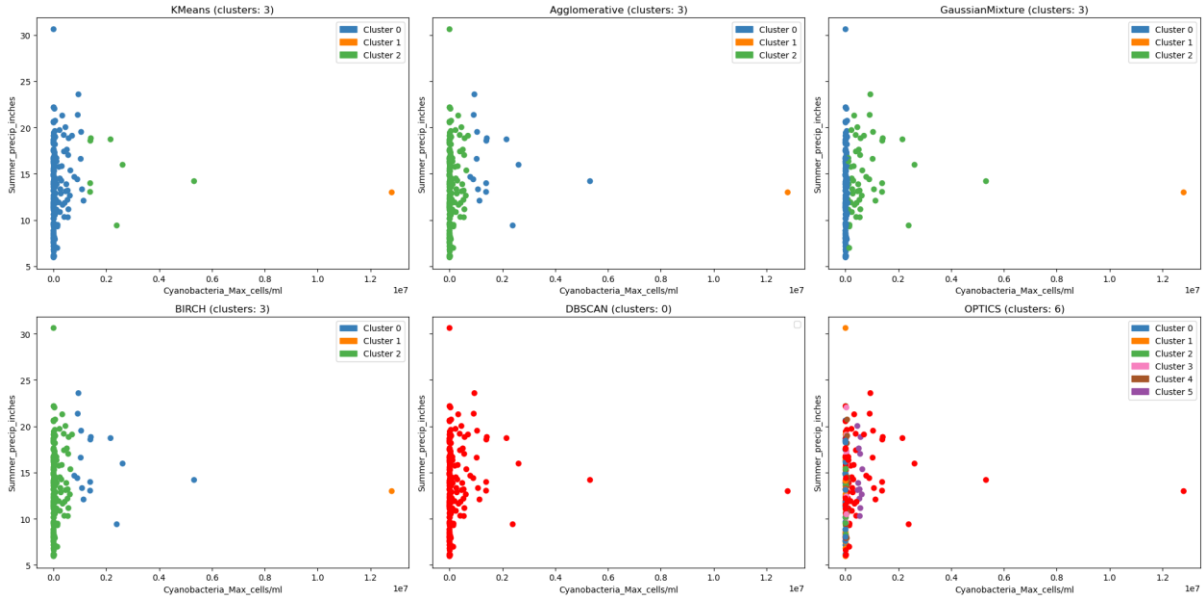
Negative Effects of Ammonia on Cyanobacteria Growth are Toxicity, Metabolic disruptions, Alteration of algal communities.

Controlling ammonia inputs to aquatic ecosystems is essential for preventing excessive ammonia levels and harmful cyanobacterial blooms. Strategies for managing ammonia inputs include:

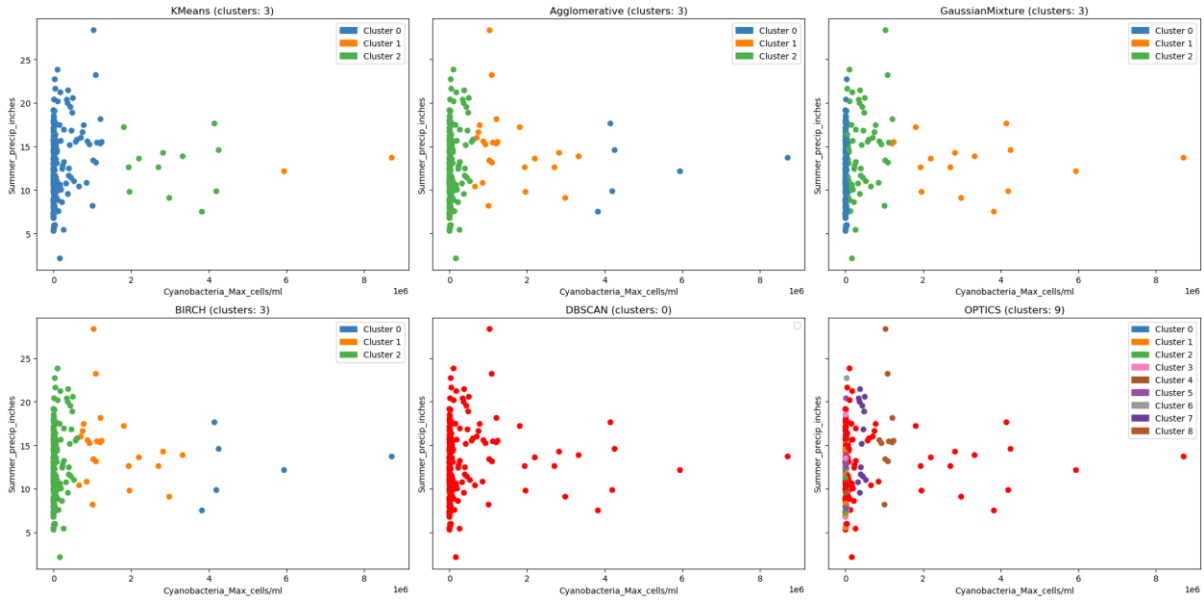
- Reducing wastewater discharge
- Controlling agricultural runoff
- Preventing urban runoff contamination

Cyanobacteria growth (cells/ml) vs Summer Precipitation (inches)

For dataset 1:



For dataset 2:



As observed from the graph, when precipitation is at its extreme, that is very low or very high, then the Cyanobacteria cell concentration reduces, this implies that precipitation is conductive for Cyanobacteria growth, but excess of it might even harm them.

Positive effects of precipitation on cyanobacteria growth are Increased nutrient availability, Stratification, Reduced grazing pressure and so on.

Negative effects of precipitation on cyanobacteria growth can be Increased flushing, Reduced light availability, Increased turbulence etc.

Overall, the effects of precipitation on cyanobacteria growth depend on the specific conditions of the aquatic ecosystem and the timing and intensity of the precipitation events.

In general, moderate precipitation events may promote cyanobacteria growth, while heavy rainfall events may have mixed or negative effects.

Comparison between all the methods used

Feature	**K-means**	**BIRCH**	**Agglomerative Clustering**	**GMMs**
Data Size	Small to medium-sized datasets	Large datasets or streaming data	Small to medium-sized datasets	Any size dataset
Number of Clusters	Known	Unknown or dynamic	Unknown or predetermined	Automatically determined
Cluster Shape	Spherical and equally sized	Any shape or size	Any shape or size	Any shape or size
Data Distribution	Assumes equal variance and independence	Can capture more complex relationships	No assumptions	Can capture complex relationships
Robustness	Not as robust to noise and outliers	More robust to noise and outliers	More robust to noise and outliers	Less robust to noise and outliers
Memory Efficiency	More memory-efficient	Less memory-efficient	More memory-efficient	Less memory-efficient
Interpretability	Provides a clear partitioning of data into clusters	Provides a hierarchical structure	Provides a clear dendrogram	Provides a probabilistic view of cluster assignments
Soft vs. Hard Assignment	Hard assignment	Hard assignment	Hard assignment	Soft assignment

Usage of Density-based Clustering methods

We tried to analyse the dataset using some density-based clustering algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and OPTICS (Ordering Points To Identify the Clustering Structure) but did not obtain any satisfactory results. The possible reasons we researched were found to be:

- As we see in the dataset, the points are too close. Due to this, DBSCAN may have difficulty identifying distinct clusters. This is because the algorithm relies on the concept of density to distinguish between clusters and noise. When points are too close together, the density becomes more uniform, making it harder for DBSCAN to identify clear boundaries between clusters. DBSCAN may either merge multiple clusters into a single large cluster or classify majority of the points as noise. A smaller eps value may lead to more clusters being merged, while a larger minPts value may result in more points being labelled as noise.

- In OPTICS, this happens for the same reason, and because of the parameters we set, we got many clusters, while the error points were significantly lesser than those obtained by DBSCAN.

- Hence, we can say that the OPTICS method is a slightly improved version of the DBSCAN algorithm

CONCLUSION

Overall, K-means is a simple and efficient algorithm that is well-suited for small to medium-sized datasets with spherical and equally sized clusters. BIRCH is a more scalable and robust algorithm that is better for large datasets or streaming data with noise and outliers. Agglomerative clustering is a good choice for exploring hierarchical structures in data, and GMMs are a good choice for modelling complex data distributions with soft assignment of data points to clusters.

Algorithm	Pros	Cons
K-means	Simple, efficient, well-suited for small to medium-sized datasets with spherical and equally sized clusters	Not as robust to noise and outliers
BIRCH	Scalable, robust to noise and outliers, memory-efficient	Less memory-efficient than K-means
Agglomerative Clustering	Good for exploring hierarchical structures in data	Can be computationally expensive for large datasets
GMMs	Can model complex data distributions with soft assignment of data points to clusters	Less memory-efficient than K-means or agglomerative clustering, not as robust to noise and outliers

The graphs acquired from both the datasets and the factual study goes hand in hand, solidifying the accuracy of our analysis. The visible representations furnished by the graphs complement the insights gained from the study, reinforcing the general findings, and strengthening the conclusions drawn. This symbiotic dating between the graphs and the look at ensures that our evaluation is grounded in each quantitative and qualitative proof, improving its credibility and persuasiveness.

REFERENCES

- **[1]** Susilo, Ahmar, Rusli, Hidayat, et al. The application of K-means clustering for province clustering in Indonesia of the risk of the COVID-19 pandemic based on COVID-19 data. Qual Quant 56, 1283–1291 (2022). https://doi.org/10.1007/s11135-021-01176-w

- **[2]** M.; Kachouie, N.N, Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. Entropy 2021, 23, 759. https://doi.org/10.3390/e23060759

- **[3]**C.,Yang, H Research on K-Value Selection Method of K-Means Clustering Algorithm. J 2019, 2, 226-235. https://doi.org/10.3390/j2020016

- **[4]**David Saulpic, Chris Schwiegelshohn. A new coreset framework for clustering. In Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing (STOC 2021). Association for Computing Machinery, New York, NY, USA, 169–182. https://doi.org/10.1145/3406325.3451022

- **[5],**X.; Zhou, X.; Lei, M.; Tepsan, W.; Deng, W.,A Novel K-Means Clustering Algorithm with a Noise Algorithm for Capturing Urban Hotspots. Appl. Sci. 2021, 11, 11202. https://doi.org/10.3390/app112311202

- **[6]**Amir Taherkordi, Øystein Haugen, Frank Eliassen, Clustering objectives in wireless sensor networks: A survey and research direction analysis, Computer Networks, Volume 180,2020,107376, ISSN 1389-1286,https://doi.org/10.1016/j.comnet.2020.107376.

- **[7]** M. -S. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796.

- **[8]**K. Masalkovaitė and P. Mohapatra, "An Overview of Fairness in Clustering," in IEEE Access, vol. 9, pp. 130698-130720, 2021, doi: 10.1109/ACCESS.2021.3114099

- **[9]** C. -D. Wang, J. -S. Wu, J. -H. Lai and C. -K. Kwoh, ;"Ultra-Scalable Spectral Clustering and Ensemble Clustering," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 6, pp. 1212-1226, 1 June 2020, doi: 10.1109/TKDE.2019.2903410

- **[10]** M.J.; Nzuza, N.; Padayachee, T.; Chen, W.; Yu, J.-H.; Nelson, D.R.; Syed, K. Comprehensive Analyses of Cytochrome P450 Monooxygenases and Secondary Metabolite Biosynthetic Gene Clusters in Cyanobacteria. Int. J. Mol. Sci. 2020, 21, 656. https://doi.org/10.3390/ijms21020656