**CIND 820: Big Data Analytics Project**

*Literature Review, Data Description, and Approach*

Sviatlana Shakibaei

501213480

Dr. Tamer Abdou

February 17, 2024

**Literature Review**

Containing nearly 20% of the earth’s fresh surface water, the Great Lakes are a global treasure. The coastlines of the great lakes and St. Lawrence River stretch from beyond Thunder Bay in the west to Atlantic Ocean in the east. These waters underpin Ontario’s high quality of life.

The Government of Ontario has been taking action to protect, conserve and restore the Great Lakes with partners, individuals, and communities to support the vision of healthy Great Lakes for a stronger Ontario – Great Lakes that continue to be drinkable, swimmable, and fishable.

In 1972 the United States and Canada signed the Great Lakes Water Quality Agreement (GLWQA) to protect and restore the waters of the Great Lakes Basin Ecosystem. Since the original version, the agreement has been revised several times and was renegotiated from 2010 to 2012. Its strengths include a structure and process that place focus on strategies for restoring and protecting the ecosystem rather than prioritizing national agendas (Alliance for the Great Lakes, 2007).

Today more than 35 million people live in the Great Lakes basin in Canada and the United States. The Great Lakes are important sources of drinking water, irrigation, transportation, and recreation opportunities such as fishing, hunting, boating, and wildlife watching. The Great Lakes are a critical component of the regional economy on both sides of the border.

Despite their great size, the Great Lakes are actually very vulnerable to [pollution](https://www.nwf.org/en/Educational-Resources/Wildlife-Guide/Threats-to-Wildlife/Pollution). The amount of water entering and leaving the lakes each year is less than one percent of the total in the lakes. Persistent chemicals that enter the lakes can remain for many years, with many building up in the food web. The source of toxic pollutants includes decades of industrial waste, raw sewage overflows, runoff from cities, and mining operations. Excess nutrients that throw the ecosystem out of balance enter the lakes from agricultural runoff and untreated sewage.

Given the spatial and temporal changes in water chemistry, regular monitoring programmes are needed to reliably estimate water quality. This leads large and complex data matrices composed of a large number of physical and chemical parameters, which are often difficult to interpret, making it challenging to draw meaningful conclusions.

Robust scientific inference is crucial to ensure evidence-based decision making. Accordingly, the selection of appropriate statistical tools and experimental designs is integral to achieve accuracy from data analytical processes. Environmental monitoring of water quality has become increasingly common and widespread as a result of technological advances, leading to an abundance of datasets. After conducting a scoping review of the water quality literature, it was found that correlation and linear regression are by far the most used statistical tools. However, the accuracy of inferences drawn from ordinary least squares (OLS) techniques depends on a set of assumptions, most prominently: (a) independence among observations, (b) normally distributed errors, (c) equal variances of errors, and (d) balanced designs. Environmental data, however, are often faced with temporal and spatial dependencies, and unbalanced designs, thus making OLS techniques not suitable to provide valid statistical inferences. Generalized least squares (GLS), linear mixed-effect models (LMMs), and generalized linear mixed-effect models (GLMMs), as well as Bayesian data analyses, have been developed to better tackle these problems. Recent progress in the development of statistical software has made these approaches more accessible and user-friendly.

An understanding of how to select the most appropriate statistical methods is critical if practitioners are to make informed decisions. When inappropriate or suboptimal methods are applied to even the most robust datasets, consequences may include drawing false conclusions, including missing environmentally critical changes.

Primary studies that focused on investigation of water quality or monitoring of water quality in ecosystems were identified by conducting a search using a pre-established list of statistical techniques and keywords. Eighteen statistical methods were identified as a basis for initial identification of studies including: Analysis of Variance (ANOVA), Bayesian Analysis, Cluster Analysis, Control Charts, Correlation, Correspondence Analysis, Factor Analysis, Kruskal–Wallis, Machine Learning, Mann-Kendall, Mann–Whitney, Generalized Linear Mixed-Effect Models (GLMMs), Linear Mixed-Effect Models (LMMs), Principal Component Analysis (PCA), Non-metric Multidimensional Scaling (NMDS), Regression, Simulation and Forecasting, and t-test.

The four most used statistical approaches included various simulation and forecasting methods (N=5,374, 34.4%), correlation (N=3,701, 23.7%), linear regression (N=1,727, 11.1%), and PCA (N=1,258, 8.1%)

**Data Description**

This section will provide exploratory analysis of the All Ontario Great Lakes dataset. All statistical analyses were conducted by using R software. Please use the following link to access a repository of the codes used to create this work: <https://github.com/SvShaki/the-All-Ontario-Great-Lakes>

The data is obtained from Ontario Data Catalogue published by the Government of Ontario on November 30, 2020. URL: <https://files.ontario.ca/moe_mapping/downloads/2Water/GLIP/All_Lakes_GLIP.csv>

As previously mentioned, the dataset includes information on sampling locations, water chemistry and chlorophyll collected at 18 locations in the Great Lakes-St. Lawrence River and 4 locations in Lake Simcoe. Data range: January 1, 1976 – December 31, 2019. Last updated: November 30, 2020.

In total the dataset contains 425011 rows and 14 attributes. The attributes, their data type and description can be found in the table below:

|  |  |  |
| --- | --- | --- |
| Attribute | Data Type | Description |
| LAKE | Character | the name of the lake where the measurement or observation was taken |
| FACILITY\_NAME | Character | the name of the facility responsible for collecting the data |
| STATION | Integer | represents a station identifier where the measurement was taken |
| DATE\_YYYYMMDD | Character | the date of the measurement in the format YYYYMMDD (Year, Month, Day) |
| YEAR | Integer | represents the year portion of the date when the measurement was taken |
| MONTH | Integer | represents the month portion of the date when the measurement was taken |
| WEEK | Integer | represents the week number corresponding to the date when the measurement was taken |
| ANALYTIC\_METHOD | Character | represents the method used for the analysis or measurement |
| TEST\_CODE | Character | represents a code associated with the specific test conducted |
| PARAMETER | Character | represents the parameter being measured or observed |
| VALUE | Numeric | represents the numeric value of the parameter measured or observed |
| UNIT | Character | represents the unit of measurement for the parameter |
| VALUE\_QUALIFIER | Character | represents a qualifier associated with the value measured |
| VALQUAL\_DESCRIPTION | Character | represents the description or explanation of the value qualifier |

The dataset initially contained 153 missing values. To ensure the integrity and accuracy of the analysis presented in this research, these missing values were addressed by removing them from the dataset prior to analysis.

The distribution of a numeric variable "VALUE" was checked by plotting a histogram of the dataset. The conclusion we have made that the distribution appears to be skewed to the right, indicating a longer right tail and a concentration of values towards the lower end. This suggests that a significant portion of observations has lower values, while a smaller proportion exhibits higher values.

Another statistical test for normality was applied - the Kolmogorov-Smirnov test. Based on the output provided, the KS test yielded a test statistic (D) of 0.4012 and an extremely small p-value (p < 2.2e-16). This indicates strong evidence against the null hypothesis that the data follows a normal distribution. Therefore, based on the KS test results, we can conclude that the variable "VALUE" in the Lakes dataset significantly deviates from a normal distribution. This implies that the assumption of normality may not be appropriate for analyses or modeling techniques that rely on this assumption. It suggests the need for alternative approaches or techniques that do not assume normality. Additionally, the warning message "ties should not be present for the Kolmogorov-Smirnov test" suggests that there might be tied values (i.e., identical values) in the dataset, which can affect the accuracy of the KS test results. It's important to address this issue if tied values are present in the dataset.

We have standardized the numeric columns using z-score standardization. Given that the data do not follow a normal distribution, standardization may be a more appropriate choice than normalization. Standardization (Z-score normalization) does not assume a specific distribution for the data and can handle non-normally distributed data effectively. It centers the data around its mean and scales it by its standard deviation, which can still be useful for algorithms and analyses that assume normally distributed data or work better with standardized features.

Normalization (Min-Max scaling), on the other hand, might not be as effective with non-normally distributed data, especially if the distribution is skewed or has outliers. Normalization could potentially amplify the effects of outliers and distort the relationships between the variables.

Therefore, based on the results of the Kolmogorov-Smirnov test indicating non-normality, standardization would likely be the better choice for preparing the data for analysis.

At the column level, some basic analysis is also offered at this point. The descriptive statistics of each column are provided.

Exploratory data analysis revealed significant insights into the "Lakes" dataset, including the distribution of sampling locations across lakes, associations between lake names and water parameters, and trends in sampling locations over time. Visualizations such as bar charts, pie charts, box plots, contingency tables, heatmaps, and time series plots were employed to elucidate patterns and trends within the dataset.

**Approach**

Analytical approach was begun by conducting the Decision Trees algorithm to assess the differences in researching pollution levels, ecosystem health, and human impacts among Lake Ontario, Lake Erie, Lake Huron, and Lake Superior. The confusion matrix and statistics obtained from the model reveal significant differences in researching between the Lakes.

A screenshot of a computer

Description automatically generated

The confusion matrix helps us understand how well our model is performing by showing the counts of correct and incorrect predictions. Reference – this is what we are trying to predict, the actual categories of the lakes. Prediction – this is what our model predicts the lakes to be. Diagonal values represent the correct predictions. For example, the number in the “Lake Erie” row and “Lake Erie” column tells us how many times our model correctly predicted Lake Erie. Off-diagonal values represent incorrect predictions. For example, the number in the “Lake Erie” row and “Lake Ontario” column tells us how many times our model mistakenly predicted Lake Ontario when it was Lake Erie. In other words, the confusion matrix shows us how often our model got each category correct and where it made mistakes.

The accuracy tells us how often the model is correct overall. In our case, the model is about 43.45% accurate, meaning it is correct a little less than half the time.

Confidence Interval (in our case CI=95%) gives us a range within which we are confident the true accuracy of the model lies. In our case, it is between 43.11% and 43.78%.

No Information Rate is the accuracy we would achieve by always predicting the majority class. In our case, it is about 39.5%.

P-Value tells us if our model is accuracy is significantly better than just randomly guessing the majority class. The p-value in our case is less than 2.2e-16, which means our model’s accuracy is significantly better than random guessing.

Kappa measures how much better our model is compared to just randomly guessing. A Kappa of 0.0798 means our model is slightly better than random guessing. In general, our model is better than guessing the majority class most of the time, but it is not incredibly accurate overall.

A chart with different colored bars

Description automatically generated

Sensitivity (True Positive Rate) tells us how good our model is at correctly identifying instances of each lake. For Lake Erie our model correctly identifies 7.34% of all instances of Lake Erie.

Specificity (True Negative Rate) shows how good our model is at correctly ruling out instances that don’t belong to each lake. For Lake Erie our model correctly identifies 98.39% of instances that are not Lake Erie.

Positive Predictive Value (Precision) measures how reliable our model is when it predicts a certain lake. When our model predicts Lake Superior, it is correct about 99.59% of the time.

Negative Predictive Value tells us how reliable our model is when it predicts that an instance does not belong to a certain lake. When our model predicts an instance is not Lake Erie, it is correct about 67.56% of the time.

Prevalence gives us the proportion of each lake in the dataset. Lake Ontario makes up about 39.50% of the instances.

Detection Rate (Recall) shows the proportion of actual instances of each lake that our model correctly identifies. Our model correctly identifies 98.53% of all instances of Lake Ontario.

Detection Prevalence is the proportion of instances our model predicts to be each lake. Our model predicts about 93.44% of instances to be Lake Ontario.

Balanced Accuracy is the average of sensitivity and specificity and gives an overall measure of how well our model performs across all classes. The balanced accuracy for Lake Erie is about 52.86%.

Precision, recall and F1 score are metrics used to evaluate the performance of a classification model, such as the decision tree model we trained to predict the lakes based on pollution levels and ecosystem health parameters.

Precision tells us how many of the instances our model classified as belonging to a particular lake actually belong to that lake. It answers the question “When our model predicts a certain lake, how often is it correct? The precision for Lake Erie is 0.69887, which means that when our model predicts an instance belongs to Lake Erie, it is correct about 69.9% of the time.

Recall (Sensitivity) tells us how many of the instances that actually belong to a particular lake were correctly classified by our model. It answers the question Out of all the instances that truly belong to a certain lake, how many did our model correctly identify? The recall for Lake Ontario is 0.98531, indicating that our model correctly identified about 98.5% of the instances belonging to Lake Ontario.

F1 Score is a measure of a model’s accuracy that considers both precision and recall. It is the harmonic mean of precision and recall, giving us a single number that represents the balance between the two. It is useful because it captures both false positives (precision) and false negatives (recall) in a single value. A high F1 score indicates both good precision and good recall. For instance, if we look at Lake Erie, the F1 score would be calculated from precision and recall values, which are 0.69887 and 0.07337 respectively.

Looking at our results:

* Lake Ontario: We have high precision (0.41650), which means when our model predicts Lake Ontario, it is correct most of the time. However, the recall (0.98531) is also high, indicating that it correctly identifies most instances belonging to Lake Ontario. This balance gives us a relatively high F1 score.
* Lake Erie: While the precision is decent (0.69887), the recall is quite low (0.07337). This means that although our model is correct when predicting Lake Erie, it misses a lot of instances that actually belong to Lake Erie. Consequently, the F1 score is lower.
* Lake Huron and Lake Superior: These lakes have similar patterns with moderate precision and recall values, resulting in moderate F1 scores.

Answering the next researching questionTop of FormAnswering the next question in our rAjejejejAhheeh, exploratory data analysis was implemented to understand the distribution of sampling locations across the lakes:

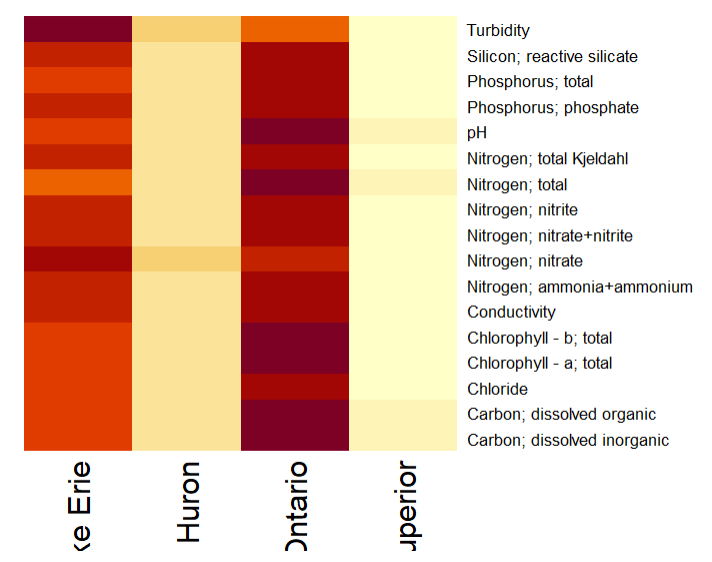
* Lake Erie 143,458 sampling locations
* Lake Huron: 68,025 sampling locations
* Lake Ontario: 167,809 sampling locations
* Lake Superior: 45,566 sampling locations

A colorful pie chart with text

Description automatically generated

Lake Ontario and Lake Erie have the highest number of sampling locations, indicating a higher density of data collection and monitoring activities compared to the other lakes. Lake Huron and Lake Superior have the fewest sampling locations among the lakes, potentially indicating lower monitoring intensity or fewer sampling programs in this region. Using visualizations such as bar charts and pie charts, we examined the number of sampling locations associated with each lake.

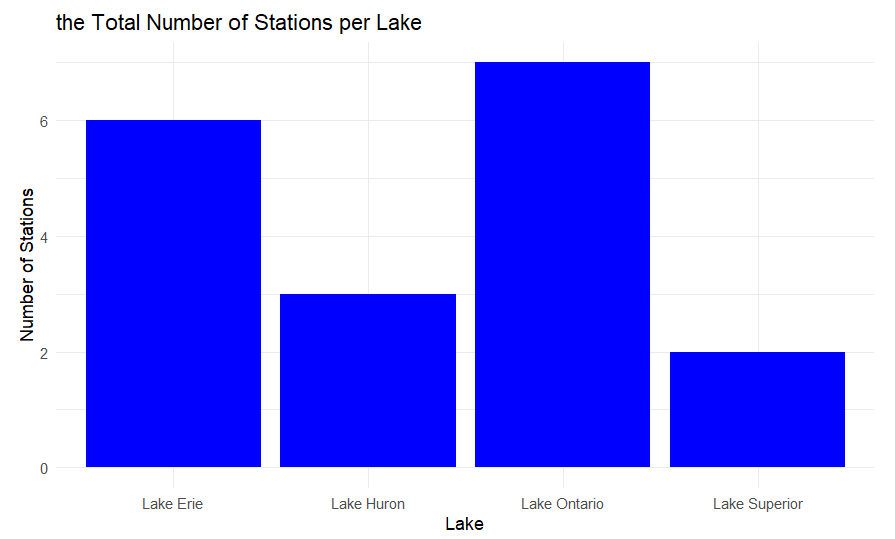
Next, the association between lake names and specific water parameters was investigated. Through contingency table analysis and calculation of mean parameter values for each lake, we identified significant relationships and variations in water quality characteristics across lakes. To visualize the association between lake name and specific water parameter a heatmap was created.



To assess the significance of differences in parameter concentrations among lakes and years, two-way ANOVA tests was performed. The analysis provided insights into variations in water quality dynamics over time and across different lake ecosystems. The analysis indicates that both the lake and year factors have a significant impact on parameter concentrations (p < 0.001 for both factors). Lake factor: The p-value associated with the lake factor is less than 0.001, indicating a highly significant difference in parameter concentrations among the lakes. Year factor: Similarly, the p-value associated with the year factor is less than 0.001, suggesting a highly significant difference in parameter concentrations across the years.

The Chi-square test of independence was conducted to analyze the association between lake name and a specific water parameter. The calculated Chi-squared statistic is 543.17 with 48 degrees of freedom. The p-value associated with the test is less than 2.2e-16, indicating an extremely significant association between lake name and the specific water parameter.

The next step in our research was the finding how many facilities associated with each lake. The Lake Ontario has the highest number of sampling locations (7). The Lake Superior has the fewest sampling locations among the lakes analyzed (2).



The primary objective of answering the next question was to identify groups of sampling locations with similar water quality characteristics. This involves finding patterns or similarities among the sampled data points to aid in targeted interventions and management strategies for water quality improvement.

Clustering techniques, such as K-Means, are widely used in data analysis and machine learning for grouping similar data points together based on their attributes or features. In our case, we are interested in grouping sampling locations based on their water quality characteristics.

K-Means is a popular clustering algorithm due to its simplicity and efficiency. It partitions the data into K clusters by iteratively assigning each data point to the nearest cluster centroid and then recalculating the centroids based on the mean of the data points assigned to each cluster. K-Means is suitable for our research because it is capable of handling large datasets efficiently and provides relatively interpretable results.

By applying K-Means clustering to our dataset, we partition the sampling locations into K clusters based on their water quality characteristics. This process enables the classification of sampling locations into distinct groups with similar patterns of water quality attributes. Each cluster represents a group of sampling locations that share common characteristics in terms of water quality. By assigning cluster labels to the original dataset, we create a basis for classification, allowing us to identify and analyze groups of sampling locations with similar water quality profiles.

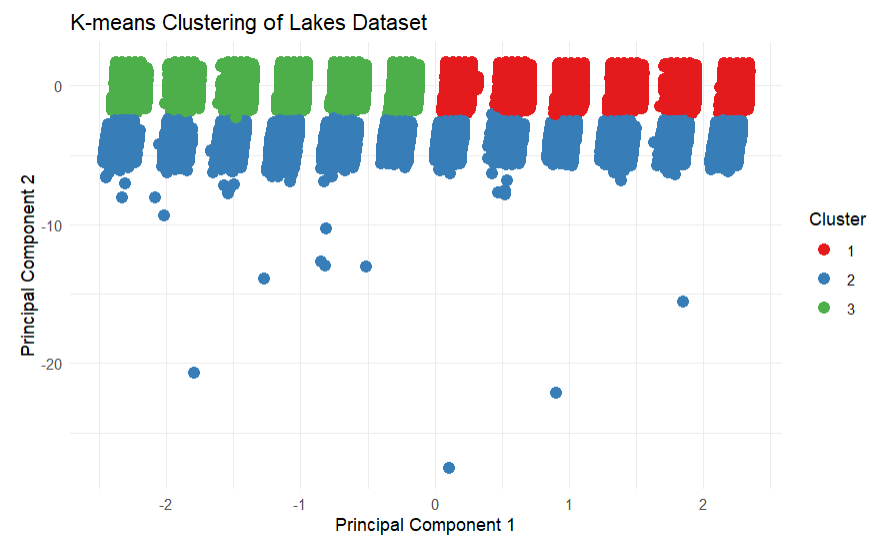
Cluster 1: This cluster exhibits average values across most variables, with slight negative deviations in STATION and VALUE.

Cluster 2:This cluster has notably high values for the variable VALUE, while other variables exhibit relatively low values.

Cluster 3: This cluster shows negative deviations in MONTH and WEEK, while other variables are close to the mean.

Cluster 1 is the largest group with 208,238 observations. Cluster 2 is the smallest group with 15,546 observations. Cluster 3 is the second largest group with 201,074 observations.

The visualization of clustering results using techniques like Principal Component Analysis (PCA) helps in understanding the structure of the data and the relationships between sampling locations. PCA reduces the dimensionality of the data while preserving its variance, making it suitable for visualizing high-dimensional datasets. By plotting the clusters obtained from K-Means clustering, we can visually inspect the grouping of sampling locations and observe any distinct patterns or trends present in the data.



By performing the Trend analysis using a Linear regression model to the time series data and analyzing the trend component, we have made the following conclusions:

* Intercept(2.347e+03): The intercept represents the estimated number of samplings locations at the beginning of the time series (YEAR = 0). In this model, it is estimated to be approximately 2347 sampling locations.
* Coefficient for YEAR (1.067e-02): The coefficient for the YEAR variable represents the estimated change in the number of sampling locations for each unit increase in the year. In this case, the coefficient is very small, indicating that there is a very slight increase in the number of sampling locations over the years.
* Significance levels: The p-value associated with the coefficient for the YEAR variable is 0.606, which is greater than the conventional significance level of 0.05. This indicates that the coefficient is not statistically significant, suggesting that there is no strong evidence to reject the null hypothesis that the coefficient is equal to zero.
* R-squared (0.001561): The R-squared value indicates the proportion of variance in the number of sampling locations that is explained by the linear regression model. In this case, the R-squared value is very low (0.001561), indicating that the linear regression model explains only a very small amount of the variance in the number of sampling locations.
* Residuals: The residuals represent the difference between the observed number of sampling locations and the number of sampling locations predicted by the linear regression model. The summary shows the minimum, 1st quartile, median, 3rd quartile, and maximum of these residuals.
* Overall, based on the results of the linear regression model, there doesn’t seem to be a significant trend in the number of sampling locations over time for each lake. The coefficient for the year variable is not statistically significant, and the R-squared value is very low, indicating that the linear model does not explain much of the variance in the data. Therefore, it seems that the number of sampling locations does not exhibit a clear linear trend over time.

Finally, for our time series analysis of the number of sampling locations over time for each lake, we have chosen to employ the AutoRegressive Integrated Moving Average(ARIMA) model. The ARIMA model is well-suited for capturing temporal dependencies and identifying underlying trends and patterns in time series data. To assess the performance of our ARIMA model, we have selected the following evaluation metrics:

* Mean Absolute Error (MAE): Represents the average magnitude of errors between the forecasted and actual values.
* Root Mean Squared Error (RMSE): Measures the square root of the average of squared differences between forecasted and actual values, penalizing larger errors more heavily.
* Mean Absolute Percentage Error (MAPE): Provides a percentage representation of the average forecast error compared to the actual values.

Based on our analysis, the ARIMA model yielded the following performance measures:

* ARIMA MAE: 136.3003
* ARIMA RMSE: 166.3437
* ARIMA MAPE: 12.86552%

These results indicate that, on average, our ARIMA model’s forecasts deviate by approximately 136.3003 units from the actual values.

The RMSE of 166.3437 suggests that the model’s forecast exhibit larger discrepancies from the actual values, particularly for outliers or extreme values.

The MAPE value of 12.86552% represents the average percentage difference between the forecasted and actual values relative to the actual values. In other words, the forecasted values differ from the actual values by approximately 12.86552%.

A graph with blue lines and red text

Description automatically generated

The plot shows how well the ARIMA model’s forecasts match the actual values over time. The blue line represents the actual values, while the red line represents the forecasted values. Ideally, these lines would be very close to each other, indicating an accurate forecast.

In conclusion, our approach to analyzing the “Lakes” dataset involved a comprehensive exploration of water quality dynamics in lake ecosystems. Through a combination of machine-learning approaches, statistical tools, techniques, and visualization, we identified significant patterns, associations, and trends in water quality parameters across different lakes and sampling locations.

**References**

1. Prachi Vasistha, Rajiv Ganguly, Water quality assessment of natural lakes and its importance: An overview, Materials Today: Proceedings, Volume 32, Part 4, 2020, Pages 544-552, ISSN 2214-7853, <https://doi.org/10.1016/j.matpr.2020.02.092>. (<https://www.sciencedirect.com/science/article/pii/S2214785320308373>)
2. Liu J, Zhang D, Tang Q, Xu H, Huang S, et al. (2021) Water quality assessment and source identification of the Shuangji River (China) using multivariate statistical methods. PLOS ONE 16(1): e0245525. <https://doi.org/10.1371/journal.pone.0245525>
3. Stefan G. Schreiber; Sanja Schreiber; Rajiv N. Tanna; David R. Roberts; Tim J. Arciszewski, Statistical tools for water quality assessment – a scoping review and recommendations for data analysis, Water Quality Research Journal (2022) 57 (1): 40–57. <https://doi.org/10.2166/wqrj.2022.028>
4. <https://www.ontario.ca/page/ontarios-great-lakes-strategy#section-15>
5. <https://data.ontario.ca/dataset/lake-water-quality-at-drinking-water-intakes/resource/a4a34b90-00f8-4e89-9519-f8ae0591fe75>
6. <https://www.ontario.ca/page/summary-canada-ontario-great-lakes-agreement>
7. <https://www.ontario.ca/document/lakeshore-capacity-assessment-handbook-protecting-water-quality-inland-lakes/monitoring-lake-water-quality>
8. Sommer Abdel-Fattah, Gail Krantzberg, A review: Building the resilience of Great Lakes beneficial uses to climate change, Sustainability of Water Quality and Ecology,

Volumes 3–4, 2014, Pages 3-13, ISSN 2212-6139. <https://doi.org/10.1016/j.swaqe.2014.11.006>

1. <https://www.nwf.org/Educational-Resources/Wildlife-Guide/Wild-Places/Great-Lakes>