**CIND 820: Big Data Analytics Project**

*Literature Review, Data Description, and Approach*

Sviatlana Shakibaei

501213480

Dr. Tamer Abdou

February 17, 2024

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%)

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. To prepare the dataset for analysis, we standardized the numeric columns using z-score standardization.

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.

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.

Analytical approach was begun by conducting exploratory data analysis to understand the distribution of sampling locations across lakes. Using visualizations such as bar charts and pie charts, we examined the number of sampling locations associated with each lake, revealing insights into spatial patterns and disparities in sampling efforts.

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 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.

Utilizing k-means clustering techniques, we identified groups of sampling locations with similar water quality characteristics. The clustering analysis enabled the classification of sampling locations based on shared patterns, facilitating targeted interventions and management strategies.

Finally, we conducted time series analysis to detect trends and patterns in the number of sampling locations over time for each lake. By visualizing temporal trends and fitting regression and ARIMA models, we gained insights into long-term variations and potential drivers of sampling location dynamics.

In conclusion, our approach to analyzing the “Lakes” dataset involved a comprehensive exploration of water quality dynamics in lake ecosystems. Through a combination of statistical techniques, clustering methods, and time series analysis, 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>