Does Water Quality Change Over Time?

A Statistical and Graphical Analysis for Minnesota Streams and Lakes at Three Different Scales

Jing Yang

Advisor: Dr. Aaron Rendahl
Acknowledgements

First I would like to thank my advisor Dr. Aaron Rendahl for giving me the opportunity to participate in this project. Thank you for all of your time, guidance, insightful feedback, and brilliant editing during these two years. I would also like to thank Dr. Yang Da and Dr. Charles Geyer for serving on my committee and for their thoughtful suggestions. I am also very thankful to people in Minnesota Pollution Control Agency. Thanks to Dana Vanderbosch for her support in coordinating my work with others in the project. Special thanks to Johanna and Laurie, who gave me much information in data and other useful material. Finally, I want to thank Li Ma for his continued support. Thanks to everyone who provided help.

This project is supported by funding from the Minnesota Pollution Control Agency.
Table of Contents

Introduction ........................................................................................................... 1

Data ....................................................................................................................... 2
  * Data Collection ............................................................................................... 2
  * Data Cleaning ................................................................................................. 5

Methods ............................................................................................................... 8
  * Trend Analysis ............................................................................................... 8
    Kendall’s Tau .................................................................................................. 8
    Seasonal Kendall Test .................................................................................... 11
    Sen Slope Estimate ......................................................................................... 13
    Permutation Test using Seasonal Kendall Test ............................................ 14
    Stouffer Z-score Test ....................................................................................... 14
  * Graphical Methods ......................................................................................... 17

Results & Analysis ............................................................................................... 18
  * Lake Station level Analysis .......................................................................... 19
  * Lake HUC level- watershed level Analysis .................................................. 25
  * Lake ECO level- Eco-region level Analysis ................................................... 28
  * Stream site level Analysis ........................................................................... 32
  * Stream HUC level-watershed level Analysis ................................................. 37
  * Stream Basin level Analysis ......................................................................... 38

Conclusion & Discussion ..................................................................................... 42

References .......................................................................................................... 44
Introduction

Minnesota is known as the “Land of 10,000 Lakes,” and there also are over 92,000 miles of streams and rivers. Because these water areas play an important role in Minnesota’s history, economy, and way of life, it is imperative that people try to monitor, maintain, and improve the quality of water in the state’s lakes, streams, and rivers. One measure of water quality is the transparency, or clarity, of the water. In water with high transparency, light reaches farther down into the water column than in water with low transparency. Low clarity in a lake or stream interferes with both plant and animal health and recreational use.

The Minnesota Pollution Control Agency (MPCA) Citizen Lake and Stream Monitoring Programs coordinate volunteers to collect transparency data on lakes and streams across the state. The Citizen Lake Monitoring Program (CLMP) has been in existence for more than 30 years. The Citizen Stream Monitoring Program (CSMP) began a little over ten years ago. These two programs combine the knowledge and commitment of interested citizens with technical expertise and resources of the MPCA. The programs greatly multiply the MPCA’s water-quality sampling capabilities, while volunteers learn more about the water quality of lakes or streams in their region and the causes and effects of water pollution.

The MPCA would like to analyze the stream or lake data to discern whether and how the clarity of water changes over time at three different scales. The scales for streams are individual stream sites, major watersheds, and major river basins. For lakes, they are lake stations, major watersheds, and eco-regions. The CLMP has previously used Systat to conduct yearly trend analysis on lake data, but the software was not easily
automated, and each year there were many problems. A year ago, a graduate student in our department developed an appropriate statistical trend analysis of CSMP data over time at the three spatial scales, and some parametric models were used for the trend analysis. Although these methods and results were promising, the MPCA realized that the analysis products required modifications and enhancements to fully suit their needs, including the investigation of stage effect for stream data, season effect for both lakes and streams data, data gaps, and censored data. This analysis addresses those concerns.

After accomplishing the task, the MPCA will have a better understanding of the water quality of the streams and lakes, and can provide the results to the volunteers to promote shared responsibility for protection of Minnesota's water resources.

**Data**

**Data Collection**

1. In the lake program (CLMP), the volunteers measure water transparency by using a tool named a “Secchi disk,” which is a circular disk used to measure water transparency in oceans and lakes. The disc is mounted on a pole or line. The following figure shows us three different kinds of Secchi disks. Various sizes and shapes of disks have been used since 1865, but the most frequently used disk is the middle one which is painted in alternate black and white quadrants. In Minnesota, people commonly use the third one, an 8-inch circular, all-white metal plate attached to a calibrated rope. Before getting started, the volunteers will receive data sheets, instructions for assembling and using the disk, a copy of the annual report on water quality of lakes in their region, and current CLMP newsletters from the MPCA.
Figure: Secchi Disk Color and Style Varieties

http://www.pca.state.mn.us/artwork/water/secchidisks.gif

About once a week during the months of May through September, the volunteer boats to a designated spot on their lakes to collect transparency readings. The location should be safe, well off-shore, and in a deep part of the lake. The volunteer should continue monitoring at that one location all five months. At that spot, the volunteer lowers the disk into the water until it is no longer visible and notes that depth from the markings on the rope. The disk is then lowered a little further and then raised back up until it is just visible. The final reading on the data sheet is the average of the first and the second depth readings. The smaller reading it is, the less clear the water become, and vice versa.

2. In the stream program (CSMP), similarly, the volunteers begin the observations with the selection of an appropriate and representative stream site. The volunteers collect both quantitative and qualitative observations for the water clarity of the stream.

The quantitative measurement is made with a transparency tube, which was developed in Australia for measuring stream water clarity. The transparency tube measures clarity in waters where a Secchi disk is not practical. This is a clear plastic tube with a release valve at the bottom. The 1.5 inch wide tube is marked in centimeters from 0 to 60 or more with a standard Secchi disk pattern at the bottom of the tube, so that this
distinct symbol is visible when looking into the tube. To use this tool, the volunteers first fill the tube with water from the stream site, look down into the tube, and release water through the valve until the Secchi pattern is clearly visible. The height of the water remaining in the tube is then recorded. If the symbol is visible when the tube is full, the transparency reading is “>60 centimeters.” The volunteers should take two transparency readings and record the average of them. As in the lake program, a greater transparency reading reflects higher water clarity.

The MPCA provides two different lengths of transparency tubes: a standard tube of 60 cm, and a longer tube of 100 cm; stream transparency data can be collected by volunteers using one or both of them.
The second way of collecting information from stream water is by qualitative visual observation. Several measurements are taken, but the only measurement used in this analysis, at the time of transparency reading, will be the estimate of the height or elevation of the stream’s water surface, also called the “stream stage.” This estimate is limited to the levels low, normal, and high, and specific guidelines are given for these levels; for example, low level means water covers one third or less of the distance from the stream bottom to the top of the bank. This is important to include in the analysis because changes in the water level, especially those due to rainfall events, can affect transparency and appearance. Other measurements include the appearance of the stream color, which is a potential cause of low transparency readings (e.g., sediment, algae, bog stain), the recreational suitability (good to poor), and the presence of a recent rainfall event. All of the details for these measurements can be found in the instruction manual provided by the MPCA.

*Data Cleaning*

1. Lake data

The original data provided by MPCA contains 414,541 individual samples on 3752 lakes between the dates of May 24, 1938, and December 4, 2009. The lakes are located in 65 watersheds and 7 larger areas called eco-regions. There are no missing values in lake data file.

To investigate the trend over years, the original dates were transformed from character to date format in R, and a new variable containing only the year of the measurement was created. To investigate the effect of seasonality, a season variable was
created according to the following table, as the date of each season depends on geographical position.

<table>
<thead>
<tr>
<th>Eco-regions</th>
<th>North</th>
<th>South</th>
</tr>
</thead>
<tbody>
<tr>
<td>Season</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Spring</td>
<td>May 1 – June 30</td>
<td>May 1 – June 15</td>
</tr>
<tr>
<td>Summer</td>
<td>July 1 – Aug 31</td>
<td>June 16 – Aug 31</td>
</tr>
<tr>
<td>Fall</td>
<td>September</td>
<td>September</td>
</tr>
<tr>
<td>Winter</td>
<td>Others</td>
<td>Others</td>
</tr>
</tbody>
</table>

There were few samples in winter and the MPCA is not interested in the lake clarity trend in winter, so the winter readings were deleted from the original data set to obtain the final data set of 390,291 observations.

2. **Stream data**

The original data provided by MPCA consists of 174,125 individual samples on 3538 stream sites between the dates of September 10, 1987, and January 21, 2010. The stream sites are located in 78 watersheds, and 10 larger areas called basins.

Because of the two tube lengths and the potential for censoring, there are four columns for recording the clarity. Two columns supply the numerical clarity measurements from the two different transparency tubes: the standard tube of 60 cm, and the longer tube of 100 cm. The other two columns supply the “>” sign for censored data if needed. These two numerical measurements were combined based on the following rules.

- If only one tube was used, keep the original measurements and modify the records from “> 60” and “>100” to “-60” and “-100,” respectively. The negative sign is used to differentiate the censored measurements from the exact measurements.
• If there are two values at the same row, the process is more complex.

  1) If both are truncated data, the final clarity measurement is defined as the larger one with a minus sign.

  2) If only the standard tube measurement is censored, the final clarity measurement is defined as the reading from the 100 cm tube if this value is equal to or greater than 60. For example, for measurements of “-60” and “80,” the final clarity measurement would be 80 cm.

  3) If both measurements are uncensored, and the absolute difference between them is less than or equal to 20 cm, the final clarity measurement is defined as the average of the two readings. For example, for measurements of “10” and “20,” the final clarity measurement would be “15.”

• Other combinations of results, such as an absolute difference of more than 20 centimeters, were treated as errors in the measurement data and deleted from the data set. For example, “20” from the standard tube and “>100” from 100 cm tube would be treated as an error.

As with the lake data, to investigate the trend over time, the original dates were transformed from character to date format, and a new variable containing only the year of the measurement was created. To investigate the effect of seasonality, a season variable was created according to the following table.

<table>
<thead>
<tr>
<th>Season code</th>
<th>Month</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>April, May</td>
</tr>
<tr>
<td>2</td>
<td>June, July</td>
</tr>
<tr>
<td>3</td>
<td>August, September</td>
</tr>
<tr>
<td>4</td>
<td>October, November</td>
</tr>
<tr>
<td>5</td>
<td>December, January, February and March</td>
</tr>
</tbody>
</table>
The final stream data set has a total of 160,728 observations through all seasons.

Since the monitoring of lakes and streams varies by volunteers, the amount of information collected from each lake station or stream site is very different. For example, there are some lake stations which just have very few observations across more than 30 years. And for stream data, the recordings of stage (low, normal and high) are not always available or correctly noted. The scarce information is a challenge to understanding the station’s clarity trend over time very well.

Methods

*Trend Analysis

The Seasonal Kendall Test was used to test for long term trends in water clarity for each lake station and stream site. This nonparametric trend test is a common method for analysis of water resources data that has been used in the field before. It compares the correlations between points at separate time periods or seasons and determines if there is a trend. This test was directly applied at the lake station level. To deal with stage effect and also censored data in stream sites, a permutation test was performed to calculate the p value. The Sen Slope estimator was used to measure the magnitude of the trend, and Stouffer's Z-score method was used to combine results across the larger spatial scales.

Kendall’s Tau

The Seasonal Kendall Test is based on Kendall’s Tau ($\tau$), which measures the strength of the monotonic relationship between $x$ and $y$. Suppose there are $n$ data pairs. The test statistics $S$ measures the monotonic dependence of $y$ on $x$. Kendall’s $S$ is
calculated by subtracting the number of “discordant pairs” $M$, the number of $(x, y)$ pairs where $y$ decreases as $x$ increases, from the number of “concordant pairs” $P$, the number of $(x, y)$ pairs where $y$ increases with increasing $x$, so

$$S = P - M,$$  \hspace{1cm} (1)

where $P$ = “number of pluses,” the number of times the $y$’s increase as the $x$’s increase, or the number of $y_i < y_j$ for all $i < j$, and $M$ = “number of minuses,” the number of times the $y$’s decreases as the $x$’s increases, or the number of $y_i > y_j$ for all $i < j$, for all $i = 1, \ldots (n-1)$ and $j = (i+1), \ldots n$.

As there are a total $n(n-1)/2$ possible comparisons to be made among the $n$ data pairs, Kendall’s Tau correlation coefficient is defined as

$$\tau = \frac{S}{n(n-1)/2}$$  \hspace{1cm} (2)

so the range of values of $\tau$ is from -1 to +1.

For a two-sided test, the null hypothesis is that no correlation exists between $x$ and $y$ ($\tau=0$), or that the distribution of $y$ does not depend on $x$, or $\text{Prob}(y_i < y_j) = 1/2$, for $i < j$.

For $n > 10$, the test statistic can be modified to closely approximate a standard normal distribution. The test statistic now is

$$Z_S = \begin{cases} 
\frac{S-1}{\sigma_S}, & \text{if } S > 0, \\
0, & \text{if } S = 0, \\
\frac{S+1}{\sigma_S}, & \text{if } S < 0,
\end{cases} \hspace{1cm} (3)$$
The null hypothesis is rejected at significance level $\alpha$ if $|Z_s| > Z_{crit}$ where $Z_{crit}$ is the value of the standard normal distribution with a probability of exceedance of $\alpha/2$. In the case where some of the $x$ or $y$ values are tied the formula for $\sigma_s$ must be modified. If two numbers cannot be compared, for example 70 and >60, or if they are equal, they are called ties. To compute $\tau$ when ties are present, tied values of either $x$ or $y$ produce a 0 rather than + or −. Ties do not contribute to either $P$ or $M$, so Kendall’s $S$ and $\tau$ are calculated exactly as before, but for the large sample approximation $Z_s$, $\sigma_s$ is adjusted to be

$$\sigma_s = \sqrt{\frac{n(n-1)(2n+5) - \sum_{i=1}^{n} t_i(i-1)(2i+5))}{18}},$$

(4)

where $i =$ number of data in each tied group,

$n =$ number of data, and

$t_i =$ number of ties of extent $i$.

For example, consider a data set of 13 values ($n = 13$) shown here in ascending order: <1, <1, <1, <1, 3, 3, 5, 5, 5, 5, 7, 7. There are a total of 3 tied groups in the data set. The largest one is of 5 values (tied at <1), the second one is of 4 values (tied at 5), there is no tied groups of 3, and there are 2 tied groups of 2 (at 3 and 7). For this data set $t_1 = 0$ (no ties of extent 1), $t_2 = 2$ (2 ties of extent 2), $t_3 =$ 0 (no ties of extent 3), $t_4 = 1$ (1 tie of extent 4) and $t_5 = 1$ (1 tie of extent 5). So for the example data,

$$\sigma_s = \sqrt{(13 \times 12 \times 31 - 2 \times 2 \times 1 \times 9 - 1 \times 4 \times 3 \times 13 - 1 \times 5 \times 4 \times 15)/18}.\]
Seasonal Kendall Test

There are many instances where changes between different seasons of the year are a major source of variation in the clarity variable. For example, stream flow always varies greatly between seasons because of sunlight, rain, snow, etc, and this variation in stream flow in turn affects water quality. To deal with seasonality, there is a fully nonparametric method based on the above Kendall test, the Seasonal Kendall test (Hirsch et al., 1982). This test accounts for seasonality by computing the Kendall test on each season separately, and then combining the results. So for monthly “seasons,” January data are compared only with January, February only with February, and so forth. No comparisons are made across season boundaries. Kendall’s S statistic $S_i$ for each season are summed to form the overall statistic $S_{all}$. Suppose we have m seasons, so the overall statistic $S_{all}$ is

$$S_{all} = \sum_{i=1}^{m} S_i.$$  \hspace{1cm} (5)

When the product of number of seasons and number of years is more than about 25, for instance, the data span 10 years and 4 seasons, the distribution of $S_{all}$ can be approximated quite well by a normal distribution with expectation equal to the sum of the expectation (zero) of the individual $S_i$ under the null hypothesis, and variance equal to the sum of their variances. The result is evaluated against a table of the standard normal distribution. The test statistic now is

$$Z_S = \begin{cases} \frac{S_{all} - 1}{\sigma_{sk}}, & \text{if } S_{all} > 0, \\ 0, & \text{if } S_{all} = 0, \\ \frac{S_{all} + 1}{\sigma_{sk}}, & \text{if } S_{all} < 0, \end{cases}$$  \hspace{1cm} (6)
where \( \sigma_{sk} = \sqrt{\sum_{i=1}^{m} \left( \frac{n_i}{18} \right) (n_i - 1)(2n_i + 5)} \), and 

\[ n_i = \text{number of data in the } i\text{th season.} \]

The null hypothesis is rejected at significance level \( \alpha \) if \( |Z_s| > Z_{crit} \) where \( Z_{crit} \) is the value of the standard normal distribution with a probability of exceedance of \( \alpha/2 \).

When some of the \( x \) or \( y \) values are tied the formula for \( \sigma_{sk} \) must be modified as above.

Compared to some parametric methods, like mixed models, these statistics are not as dependent on assumptions about data distribution. The Seasonal Kendall test is also highly robust and relatively powerful, and can handle censored data. It seems a better way to handle our data.

To use the Seasonal Kendall test, at least two years are needed. However, for the lake or stream with sample size less than 6, we determined that there is not enough information to tell us about the water clarity trend so we did not analyze them. As I mentioned earlier, for lake data, there are three seasons, spring, summer, and fall, so at the lake station level, the test is exactly the Seasonal Kendall test by performing a Kendall test in each season and combining three test results. For analysis at stream site level, there are five seasons according to the suggestion from MPCA. Since the stage effect was also needed to be taken into account in steam site level analysis, the test was different depending on the number of observations for each stream site. If one stream site included more than five observations with the stage values, the data in this stream site were compared within fifteen season and stage combinations due to the five seasons and three stage levels, which means we have fifteen “seasons” here. For the rest of stream sites with sample size larger than five, the data were just compared within five seasons.
**Sen Slope Estimate**

We also calculated the Sen Slope and intercept estimates to measure the magnitude of the trend.

At Lake station and Stream site level, the Sen Slope estimate is the median of all two point slopes over years. These slopes are computed within each season. Suppose the data is \((X_{ij}, Y_{ij})\), where \(X_{ij}\) denotes the \(i\)th year in season \(j\) and \(Y_{ij}\) denotes the median of all clarity records in the \(i\)th year and \(j\)th season, for season \(j = 1, 2, \ldots, p\). Then the Sen Slope estimate is

\[
\hat{\beta}_1 = \text{Median}_{k < i} \left( \frac{Y_{kj} - Y_{ij}}{X_{kj} - X_{ij}} \right),
\]

\[j = 1, 2, \ldots, p\]  \hspace{1cm} (7)

and the overall estimate of intercept is the median of the \(p\) season estimate of intercept, so

\[
\hat{\beta}_0 = \text{Median}(\hat{\beta}_{0,1}, \hat{\beta}_{0,2}, \ldots, \hat{\beta}_{0,p}) \quad \text{and}
\]

\[
\hat{\beta}_{0j} = Y_{0.5j} - \hat{\beta}_{1j}X_{0.5j},
\]

where \(X_{0.5j}\) and \(Y_{0.5j}\) denote the sample medians of the \(X\)'s and \(Y\)'s, respectively, for season \(j\), and \(\hat{\beta}_{1j}\) is the estimate of slope in each season. The slope and intercept at the larger levels are also calculated by averaging all of the slopes or intercepts from the tests being combined, respectively.

At larger region levels, the overall estimates of slope and intercept are the mean values of all slopes or intercepts within the current region, respectively.
Permutation Test using the Seasonal Kendall Test

For most of hypothesis tests, we usually start with the assumptions and work forward to derive the sampling distribution of the test statistic under the null hypothesis. As I mentioned, the sampling distribution of the Kendall test for large sample under our null hypothesis is closely approximated by a normal distribution. At the stream site level, we have two kinds of censored values, “>60” and “>100,” and also some values between 60 and 100, like 70, 85 etc. We can’t hold the normal approximation assumption anymore, because we do not have groups of ties in the same way as I introduced above. To deal with this kind of censored data, we perform the following permutation test for testing the hypothesis that no correlation exists between \( x \) and \( y \).

For each “season,” we calculate the true Kendall \( S_t \), and the true sum \( TS \) is

\[
TS = \sum_{i=1}^{m} S_t .
\]  

(10)

We then randomly rearrange the transparency data \( (Y) \) by holding ascending time data \( (X) \) within each “season” and calculate a new Kendall \( S_t \), denoted \( Sperm_t \) to distinguish it from true Kendall \( S_t \) value, and the permuted sum is

\[
PS_j = \sum_{i=1}^{m} Sperm_t .
\]  

(11)

We repeated this procedure 1000 times to obtain \( PS_1, PS_2, \ldots, PS_{1000} \), and the P value compute as

\[
Pr(|PS|>|TS|) = \frac{\text{total number of } |PS_t|>|TS|, \ t=1,2,\ldots,1000}{1000} .
\]  

(12)

Stouffer’s Z-score Test

At the lake station and the stream site level, we can test whether the particular lake station or stream site’s clarity is changing over time. It would be difficult to get an
overall feel of how the state’s water clarity is doing by just looking at a particular lake or stream. For example, the southern part of the state might be very different than the northern part of the state as a result of climate or different environment, etc. To get a better understanding of the state’s clarity, the MPCA asked for an analysis at two larger regional levels, (1) lake and stream watersheds, and (2) lake eco-regions and stream basins.

Stouffer's Z-score method was used for the analysis on these larger levels. This technique for data fusion is named for the sociologist Samuel A. Stouffer, and its basic form is used to combine the results from several independent tests bearing upon the same overall hypothesis. It first converts the p-values from each test into Z-scores,

\[
Z_i = \left| \Phi^{-1} \left( \frac{p_i}{2} \right) \right| \cdot \text{sign}(Z_i),
\]

(13)

where \( \Phi \) is the standard normal cumulative distribution function, and \( p_i \) is the p-value for the \( i^{th} \) hypothesis test, and combines them to obtain a Z-score for the overall meta-analysis,

\[
Z = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}},
\]

(14)

where \( k \) is the number of independent tests being combined and \( w_i \) is the weight for the \( i^{th} \) Z-score. \( Z \) has a standard normal distribution, so a two-sided p-value

\[
P = 2\Phi(-|Z|),
\]

(15)

which can be determined for \( Z \) easily.

Stouffer's Z-score method is typically applied to a collection of independent test statistics, usually from separate studies having the same null hypothesis. The null hypothesis is that all of the separate null hypotheses are true. The alternative hypothesis
is that at least one of the separate alternative hypotheses is true. In the case of a meta-
analysis using two-sided tests, the null hypothesis is usually the same as using one-sided
tests, but the alternative hypothesis does not specify a particular effect direction. It is
possible to reject the overall meta-analysis null hypothesis even when the individual
studies show strong effects in different directions. In this case, we will reject the
hypothesis that the null hypothesis is true in every test, but this does not imply that there
is a uniform alternative hypothesis that holds across all studies. Therefore, two-sided
meta-analysis is particularly sensitive to heterogeneity in the alternative hypotheses. In
our analysis, we performed two-sided tests, and the separate null hypothesis is that no
correlation exists between time and clarity.

All of the tests at lake station or stream site level are considered to be independent
because the measurement locations are comparatively distant, and the sampling of
streams and lakes was carried out independently. In the case that the tests are not
independent, dependence among the statistical tests are generally positive, which means
that the overall p-value will be too small if the dependency is not taken into account.
Therefore, if Stouffer’s Z-score method for independent tests is applied in a dependent
situation and the p-value is not small enough to reject the null hypothesis, then that
conclusion should continue to be held even if the dependence is not properly accounted
for. However, if the positive dependence is not considered and the p-value is too small,
there might be the evidence of overstating the alternative hypothesis. In future analysis,
one could consider this issue to get more in-depth results.

We treat every $Z_i$ as having equal weight in our analysis.
**Graphical Methods**

Graphs are often a helpful way to check and display our results. It provides a visual clue for water clarity change over time. The smoothing curve is the major method used for our graphical analysis.

Since there are season and stage effects, we decided to fit smoothing models by using additive models at lake station and stream site level. The additive models were performed by *gam* function from *mgcv* package in R. At the larger regional levels, because the samples from the same lake or stream are not independent, we took this into account and extended the *gam* models to allow for correlation between the observations in the same lake or stream by using *gamm* function where lake or stream were treated as random effects.

The additive model is a nonparametric regression method with the following description.

Suppose we have a data set \( \{y_i, x_{i1}, \ldots, x_{ip}\}_{i=1}^n \) of n statistical units, where \( x_{ij} \) \((i=1,\ldots, n \text{ and } j=1,\ldots, p)\) represent predictors and \( y_i \) is the response. The additive model takes the form

\[
Y = \beta_0 + \sum_{j=1}^{p} f_j(X_j) + \varepsilon,
\]

where \( E[\varepsilon]=0, \text{Var}(\varepsilon)=\sigma^2 \) and \( E[f_j(X_j)]=0 \), and the functions \( f_j(x_{ij}) \) are some smooth functions fit from the data.

So, for our data, the additive models are,

At lake station level, \( Y_i = \beta_0 + f_1(X_i) + f_{\text{factor}(\text{season}_i)} + \varepsilon_i \)

At stream site level, \( Y_i = \beta_0 + f_1(X_i) + f_{\text{factor}(\text{season}_i)} + f_{\text{factor}(\text{stage}_i)} + \varepsilon_i \)

At larger regional levels, random variable such as lake or stream was added in the model.
Here, the functions $f_1(x_i)$ may be fitted by using parametric or nonparametric means. At this point, the nonparametric fits could make relaxed assumptions comparing with the parametric fits, and perhaps make the interpretation of results easier to someone who does not need all the theory and background of how it works.

For stream site, since we have many truncated data point, I decided to impute some data points in order to make a better plot. Before fitting an additive model, when there are some non-truncated clarity values larger than 60 or truncated “$>100”$ values, I substituted all of truncated data points having values “$>60”$, with the median of all absolute values larger than 60.

At the lake ECO-region level and stream Basin level, because sample size is too large to be handled by fitting additive model in R. We decided to use the smoothing spline directly for producing the figures. And we also adopted the new values calculated by the following formula in smooth spline for removing the lake or stream effect in the same region.

$$\text{New value} = O + M - m,$$

where $O$ denotes the original clarity values of every lake or stream. $M$ denotes the overall mean of all of lake/stream clarity values in one ECO/Basin. And $m$ is the mean value of all original values in one lake/stream.

**Results & Analysis**

All of the analyses of CLMP and CSMP data were implemented by using the statistical software R, version 2.11.0.
Lake station level Analysis

The lake stations varied by the amount of data collected. Some lake stations have consistent data across many years while other lake stations just have a couple of observations in one year or several years. For the sake of better understanding, three different colors were used for showing the different seasons in our plot. We denote spring by green circle dots, red color is for summer, and blue circle dots stand for fall. Here is a reminder that the fitted lines plotted come from additive models, which gives us a nonparametric fitted line.

As we known, the original data contains 3,752 lake stations. The available lake stations for Kendall test are 2,331.

Among of them, some lake stations have lots of observations. For example, lake named Minnetonka (upper lake) with station ID 27-0133-05 in the central Minnesota (Figure 1), has about 2000 observations through about 40 years, which means the volunteers have been consistent in the monitoring of this lake station from 1969 to 2009. The figure shows us the data points and the clarity trend over time. Looking at the dots in different seasons, the lake in the spring is more clear than in the summer and fall, and the similar situations happened in the most of lake stations. Looking at the overall picture of across the years, clarity is getting a little better, just as the test result told us. The overall Kendall’s $S$ is 41596, which means there are about 40,000 more concordant clarity and years pairs than discordant pairs through 40 years, and its standard deviation $\sigma_{sk}$ is about 10440.31. After standardizing the overall $S$ by subtracting its expectation $\mu = 0$ and dividing by $\sigma_{sk}$, the test statistic $Z_s$ is about 3.98 which is larger than 1.97, which means we can reject the null hypothesis that no correlation exists between water clarity and time.
The corresponding p-value is about 7e-05 as shown on the following figure. The positive evidence also can be drawn from the value of the Sen Slope, say 0.027, which gives us the median of all two point slopes computed within each season.

![Graph showing water clarity over time for Minnetonka (Upper Lake), MN](image)

**Figure 1. Minnetonka (Upper Lake), MN**

Some of lake stations with modest amount of samples still can give evidence for water clarity such as Wilkins lake in Aitkin County, MN (Figure 2). There just 327 observations for this lake station, starting in 1977 - 2009. But there seems to be enough data to analyze this lake clarity. The overall Kendall’s $S$ is 10753, and its standard deviation $\sigma_{sk}$ is about 934.8. So the sample approximation $Z_s$ is about 11.05 which is large enough to reject the null hypothesis that no correlation exists between water clarity.
and time. The corresponding p-value is about 2.2e-28. The overall water clarity trend over time is obviously getting better from the figure 2, in which the clarity has a little bit high values in the spring comparing with values in the summer and fall. The positive evidence also can be drawn from the value of the Sen Slope which is 0.046.

Figure 2. Wilkins Lake in Aitkin County, MN

Figure 3 is a lake station with a significant decreasing clarity trend over time. This site is located near Stillwater, Minnesota. This lake station had a clear decreasing significant clarity trend with a Sen Slope -0.078. The p-value is about 1.88e-34 which is rounded to zero on the following figure.
Figure 3. Square Lake near Stillwater, MN

Figure 4 shows us one lake station with no clarity change over years. The overall Kendall’s $S$ is -731, which means there are 731 more discordant clarity and years pairs than concordant pairs through 25 years, and its standard deviation $\sigma_{sk}$ is about 513.46. So the test statistic $Z_s$ is about -1.42 which means we have lack of evidence that there is a correlation between water clarity and time. The corresponding p-value is about 0.155 as shown on the following figure. By looking at the graph, it appears that we didn’t find obvious clarity trend most of time. **Although there is a large dip in recent years, we still got non-significant result, because we have only 11 data points after 2005, and there is no**
decease trend before 2005. In the future, we could know the trend better by having more records from this lake.

Figure 4. Sugar Lake in Wright County, MN

Definitely, data is not always perfect for analysis. Of course, some lakes have very few observations. With few data points on hand, it is hard to perform a right analysis. For instance, a lake station called Rock lake, is located in north of Pillager, MN (Figure 3). There are only 7 observations for years 1981 to 2009, with most of years having no record. One thing we can take away from this graph is the clarity readings are very low in two years, 1981 and 2009. The values are just about 1 meter unlike the clarity values of the lakes above. Taking infrequent and inadequate data collections is not advisable,
because it is hard to perform an analysis without a lot of data on hand. In this kind of situation, we will not be able to know the water clarity very well.

Figure 5. Rock lake in north of Pillager, MN

Provided those graphs are just a snapshot of the 3752 lake stations and the three different scenarios that could happen at each lake station: positive significance trend, no trend and negative significance trend. In order to get a rough sense of all the lake stations clarity trends, the following table shows us a summary according to the Seasonal Kendall test results, p-values.
Table 1. Summarization of all the trends of the lake stations

<table>
<thead>
<tr>
<th>Number of lake stations having a significant increasing trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>The slope is non-negative and has a p-value that is less than or equal to 0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of lake stations having a significant decreasing trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>The slope is negative and has a p-value that is less than or equal to 0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of lake stations having no trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>The a p-value is greater than 0.05</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of lake stations having no test result (with very few data)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
</tr>
</tbody>
</table>

*Lake HUC level- watershed level Analysis*

Recall the Stouffer's Z-score method used for testing the clarity trend at the watershed level and Eco-region level.

Then the test statistic for the overall lake HUC is

$$Z = \frac{\sum_{i=1}^{k} w_i Z_i}{\sqrt{\sum_{i=1}^{k} w_i^2}}$$  \hspace{1cm} (18)

where $k$ is the number of independent tests being combined, and $w_i$ is the weight for the $i^{th}$ Z-score. We treated every lake station as having equal weight in our analysis. So $Z$ has a standard normal distribution, so a two-sided p-value can be determined for $Z$ easily:

$$P = 2\Phi(-|Z|).$$

There are 65 HUC regions for lakes in Minnesota. Some of regions include hundreds of lake stations with clarity records, and some others just have one or a couple of lake stations with data. The more data and lake stations for one HUC region we have, the more information we could gain from them.
The HUC region with ID 7010206 (Figure 6), includes 206 lake stations. The figure shows below. The black dash lines are fitted lake stations trend lines from lake station level analysis, and the red full line is fitted clarity trend line getting from an additive model by considering season effect and the lake stations random effect. Both of the HUC trend line and the cluster of lake station trend lines display an increasing trend. This conclusion was confirmed by the p-value and slope again. The calculated meta-analysis Z-score is 12.69 and the corresponding p-value is about 6.43e-37. Recall the separate null hypothesis is that no correlation exists between time and clarity. The test results suggest that the null hypotheses are not true for every test in this HUC region. And the calculated positive Sen Slope, 0.019, which is the mean of all slopes of lakes in this HUC region, shows us there might be an increasing clarity trend over time in this HUC region.

Figure 6. Lake HUC Region (7010206), MN
Figure 7 shows us another HUC region (4010102). There are only 5 lake stations in this region. Three of them have lots of missing values in most of years, for this reason, the result should be carefully thought out. For this region, there is no significant water clarity trend in this HUC region. The corresponding p-value is 0.559 as shown on the figure.

Figure 7. Lake HUC Region (4010102), MN

Table 2 gives us the summarization of the analysis results at the lake HUC level.
### Table 2. Summarization of all the trends of the lake HUC regions

<table>
<thead>
<tr>
<th>Description</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of regions having a significant increasing trend</td>
<td></td>
</tr>
<tr>
<td>The slope is non-negative and has a p-value that is less than or equal to 0.05</td>
<td>27</td>
</tr>
<tr>
<td>Number of regions having a significant decreasing trend</td>
<td></td>
</tr>
<tr>
<td>The slope is negative and has a p-value that is less than or equal to 0.05</td>
<td>11</td>
</tr>
<tr>
<td>Number of regions having no trend</td>
<td></td>
</tr>
<tr>
<td>The p-value is greater than 0.05</td>
<td>23</td>
</tr>
<tr>
<td>Number of regions having no test result (with no lake station)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4</td>
</tr>
</tbody>
</table>

*Lake ECO level- Eco-region level Analysis*

The state of Minnesota is divided into seven eco-regions, as shown in the following map (Picture 1) and table. Eco-regions are areas of relative homogeneity that were developed from mapped information based on land surface form, soils, land use, and potential natural vegetation. Recall the smooth spline used for producing the graphs.
There are two graphs displayed and are organized by similarity in the fitted Eco-region water trend lines. In figure 8 (a), four Eco-regions have significant test results with p-values less than 0.05 and a positive calculated Sen Slope which means positive clarity...
trend over time, and for figure 8 (b), the test results indicate that there is no significant clarity change over time for each lake station in these three Eco-regions. Average over the whole period varies by region. From the figures, the “NLF” Eco-region has the highest long term average than the others. Some Eco-regions just have data for a shorter time period. For instance, for the Eco-regions “DA,” “RRV” and “NMW,” the years monitored no longer go from 1938 to 1970, and they have clarity data after 1970, so analysis were done on only these years.

(a)
To wrap up the lake Eco level analysis, here is a table depicting the results from the test of clarity over time.

**Table 3. Summarization of Lake Eco-region trends**

<table>
<thead>
<tr>
<th>Lake Eco-region</th>
<th>P vaule</th>
<th>Slope</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>DA</td>
<td>9.22E-05</td>
<td>0.043</td>
<td>Positive trend</td>
</tr>
<tr>
<td>NCHF</td>
<td>1.03E-110</td>
<td>0.013</td>
<td>Positive trend</td>
</tr>
<tr>
<td>NGP</td>
<td>0.440987675</td>
<td>0.021</td>
<td>No change</td>
</tr>
<tr>
<td>NLF</td>
<td>5.45E-86</td>
<td>0.038</td>
<td>Positive trend</td>
</tr>
<tr>
<td>NMW</td>
<td>0.196611265</td>
<td>0.103</td>
<td>No change</td>
</tr>
<tr>
<td>RRV</td>
<td>0.045158886</td>
<td>0.056</td>
<td>Positive trend</td>
</tr>
<tr>
<td>WCBP</td>
<td>0.259507326</td>
<td>0.008</td>
<td>No change</td>
</tr>
</tbody>
</table>
*Stream site level Analysis*

The process for analyzing stream data is a little bit more complicated than that for lake data. First, the stream level data are highly unbalanced: the amount of data collected from different stream sites varied due to different volunteers’ habits just very similar to lake data. Second, we need to consider one more variable, stage (low, normal, high), for stream data analysis. The other main issue is that we have a large amount of truncated data in stream data as a result of using the two kinds of transparency tubes. In order to distinguish truncated data from normal data in our presentation, we used two different circles in our figures. For example, we denote “>60cm” and “>100cm” by solid dots, and use hollow circle dots for “60cm” and “100cm,” respectively. Again, the fitted lines are calculated from additive models by using estimated smoothing curves. For getting more reasonable plot, imputation for truncated values “>60” was done before fitting models.

The original stream data contain 3,538 stream sites, but we have only 2,213 available stream sites for permutation test using the Seasonal Kendall method due to the unbalanced data. Some sites are removed because they do not have enough data for a test. For stream data, there is a new trouble about data size, which is about one third data have no stage records for total 160,728 rows. Some of stream sites can only be test without including stage effect. In order to make the best use of our data and model the stage effect reasonably, we tested some stream sites with stage effects if having enough stage records, while tested the rest of stream sites without stage effects. The following section gives some specific examples for several different situations.

A stream site with enough stage records is shown in Figure 9. The black solid dots denote the data “>60 cm” and “>100 cm.” This stream site has 343 observations through
about 10 years, which means the volunteers have been consistent in the monitoring of this lake station from 2000 to 2009. This stream site had a clearly increasing trend in clarity with a calculated Sen Slope 1.866. The p-value from the permutation test is close to be zero after formatting the value to have three decimal places in the figure.

Figure 9. Stream site “S000-230” in MN

Another example I would like to mention is a stream site that has enough observations while there is few stage records for it. For total 2,213 available stream sites, there are 884 sites having no sufficient stage records, so we have to test them without stage effect. As shown in figure 10, stream site “S002-152” has 260 observations across 8 years, but only has 2 observations with stage records. Here, in the figure, there seems to
be an increasing trend over time, especially in the last few years, but the test result tells us the water clarity does not get better or worse significantly (p-value=0.1).

![Graph showing water clarity trend](image)

**Figure 10. Stream site “S002-152” in MN**

As shown in Figure 11, stream site “S001-781” has a significant decreasing clarity trend over time. There are a total of 226 valid observations with stage values in this site. This stream site had a clear decreasing clarity trend with a calculated Sen Slope -2.75. The p-value is very small which is close to zero by formatting the value to have three decimal places in the permutation test’s outcome as shown on the following figure.
It is not easy to perform a sound statistical test without enough data. In addition, we have a lot of censored data in the stream and lake level data. Again, to handle these censored data, we took the median of all of numbers which are larger than 60 instead of every censored data points “>60” in the additive models. My speculation for this site is that there are two possible situations. One is that the water clarity is getting better or worse. For example, the exact water clarity values might be from 150 cm to 100 cm across the years, but we cannot get this information from these data. And the other one is this stream is always clear at the majority of time. A recommendation would be to use the larger, such as 120 cm or longer, transparency tube to get more variable and accurate results for this kind of stream sites.
These example graphs are only a fragment of the 3538 stream sites and these different scenarios could happen at each stream site with or without stage effect. In order to get a rough sense of all stream site clarity trends, the following table shows us a summary according to the test results and p-values.

Figure 12. Stream site “S000-007” in MN
Table 4. Summarization of all the trends of the stream sites

<table>
<thead>
<tr>
<th>Number of stream sites having a significant increasing trend</th>
<th>492</th>
</tr>
</thead>
<tbody>
<tr>
<td>The slope is non-negative and has a p-value that is less than or equal to 0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of stream sites having a significant decreasing trend</th>
<th>113</th>
</tr>
</thead>
<tbody>
<tr>
<td>The slope is negative and has a p-value that is less than or equal to 0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of stream sites having no trend</th>
<th>1608</th>
</tr>
</thead>
<tbody>
<tr>
<td>The a p-value is greater than 0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of stream sites having no test result (with very few data)</th>
<th>1325</th>
</tr>
</thead>
</table>

*Stream HUC level-watershed level Analysis*

The Stouffer's Z-score method is used for testing the clarity trend at the stream watershed level and stream basin level. I performed the same procedure as I did for lake data analysis in larger levels.

There are 78 HUC regions for streams in Minnesota. 77 of them are available to do the test, since HUC 9030003 has no clarity record. The analysis at stream HUC level and Basin level is analogous with what I discussed in lake part. Here is the summary table for the analysis results at the stream HUC level.

Table 5. Summarization of all the trends of the lake HUC regions

<table>
<thead>
<tr>
<th>Number of regions having a significant increasing trend</th>
<th>39</th>
</tr>
</thead>
<tbody>
<tr>
<td>The slope is non-negative and has a p-value that is less than or equal to 0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of regions having a significant decreasing trend</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>The slope is negative and has a p-value that is less than or equal to 0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of regions having no trend</th>
<th>36</th>
</tr>
</thead>
<tbody>
<tr>
<td>The a p-value is greater than 0.05</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Number of regions having no test result (with no stream site)</th>
<th>1</th>
</tr>
</thead>
</table>
*Stream Basin level Analysis*

The state of Minnesota is divided into ten basins. Picture 2 shows us all of the 10 basins and their locations in Minnesota. Below the picture is a table with the basin names and number of stream sites included in the analysis.

![Picture 2. Minnesota Map with the Major Basin Regions](http://www.gda.state.mn.us/maps/RiverBasins.gif)
The water quality trend was summarized in two graphs, which are organized by similarity in the fitted basin water trend lines. In figure 13(a), seven basins have significant increasing water clarity over time with positive calculated Sen slopes, and in figure 13(b), the test results indicate that there is no significant clarity change over time for each stream site in these three basins. Average over the whole period varies by region just as you can see. Three basins have continues high water quality. From the figures, the “LS,” “SC,” and “RN” basins have the highest long term average than the others.
Figure 13. Fitted lines for the stream Basins of MN
To wrap up the stream Basin level analysis, here is a table depicting the results from the test of clarity over time.

**Table 6. Summarization of Stream Basin trend**

<table>
<thead>
<tr>
<th>Stream Basin</th>
<th>P value</th>
<th>Slope</th>
<th>Trend</th>
</tr>
</thead>
<tbody>
<tr>
<td>CD</td>
<td>7.74E-05</td>
<td>3.681</td>
<td>Positive trend</td>
</tr>
<tr>
<td>DM</td>
<td>0.774115</td>
<td>1.394</td>
<td>No change</td>
</tr>
<tr>
<td>LM</td>
<td>2.72E-28</td>
<td>2.712</td>
<td>Positive trend</td>
</tr>
<tr>
<td>LS</td>
<td>0.011371</td>
<td>1.935</td>
<td>Positive trend</td>
</tr>
<tr>
<td>MN</td>
<td>1.08E-126</td>
<td>1.914</td>
<td>Positive trend</td>
</tr>
<tr>
<td>MO</td>
<td>0.246253</td>
<td>0.746</td>
<td>No change</td>
</tr>
<tr>
<td>RD</td>
<td>4.92E-12</td>
<td>2.168</td>
<td>Positive trend</td>
</tr>
<tr>
<td>RN</td>
<td>0.00155</td>
<td>3.734</td>
<td>Positive trend</td>
</tr>
<tr>
<td>SC</td>
<td>0.800601</td>
<td>1.081</td>
<td>No change</td>
</tr>
<tr>
<td>UM</td>
<td>9.34E-23</td>
<td>2.748</td>
<td>Positive trend</td>
</tr>
</tbody>
</table>
Conclusion & Discussion

Over the past dozens of years, the MPCA has monitored and managed the quality of Minnesota’s waters, with the help of volunteers, in order to discern the water quality of the lakes and streams and help protect this important resource. The objective for this project is to give feedback to the volunteers on the quality trend of their waters.

Using the methods and analysis in this report, we could provide the graphical displays and test results on how lakes and streams are doing, how the watersheds are doing, and how eco-regions or basins are doing in terms of water quality. The MPCA are able to give the volunteers both visual images and quantitative results about clarity change over time using these tools.

After passing through this project, here are some recommendations I would like to mark. For data quality consideration, using single longer transparency tube will eliminate unclear measurements and allow for more variability in the larger measurement readings. The second suggestion is to monitor more stations or sites for better understanding the quality trend at the larger levels. Because some larger regions are fairly large and only have one or a couple of sites or stations monitored, it is not enough to gain relevant information on the whole region. Also, while stage was an interesting factor, this variable is not used for many stream sites due to the lack of consistency by the volunteers in marking the qualitative visual estimate of the water level. In the further monitoring, with more stage records, the analysis will be more accurate.

For further analysis, we would suggest to consider the case that the tests in the same larger regions are not independent, and different lakes or streams take different weights in Z-score test, since there are the different numbers of the data points between
lakes or streams. And also the issue of different number of the data points between the seasons and years for each lake or stream should be a concern. I believe there should be better ways to handle it. After considering all of these points, we will avoid some possible errors and increase more power in the future analysis.
References


2. Laura (Freer) Le. Are Stream’s Clarity Getting Better Over Time? A Statistical and Graphical Analysis of MN Streams and MN River Basins


4. MPCA. Citizen Stream-Monitoring Program website

5. MPCA. Citizen Lake-Monitoring Program website