Thesis

Nick Zinck

March 2018

Contents

1 Introduction

1.1 Problem Statement

A Reservoirs water quality effects the type and extent of neccesary treatment process of the water supply prior to distribution. Generally, less treatment is required for a water utility who's source is a remote, healthy reservoir than a reservoir which has been degraded by anthroppologic or other means. It is often infeasible to obtain water from completely pristine areas, especially for water utilities who supply water to urban areas. Watershed management can help ensure the water quality of a reservoir and as well as predict, lessen, or prevent reservoir water quality degredation. from occuring. (Quantity as well.)

Water Quality data collection and analysis can assist on decision making for utilities. It is not feasible to know every, yet an adequate sampling plan can bring insight about the condition of the reservoir as well as the tributaries entering the reservoir. It is not just enough to collect this data, the task fo water quality monitoring does not stop after the field or after the lab. Water Qulaity data analysi scan help answer questions and suggest solutions once problems arise. Even in the case of non problemtic situations water qualit data can help one understand the unique processes of the reservoir in question, calibrate water quality models, and suggest the updated sampling plans. Questions must be asked like is the current sampling plan adequate to get as good of an undestanding of the water quality processes that are happeing in the reservoir and watershed, within reasonable time and cost constraints.

1.2 Objective

A comprehensive watershed monitoring program includes collecting much water quality, meteorological, and hydrological data. Although data collection is always important, much of this data is underutilized due to a timely processes of searching, displaying, and analyzing data. Spreadsheet applications are likely not an effective way to store large datasets and visualization and analysis tools are limited. Relational database applications serve as a better home for these large datasets, yet visualization and analysis tools are commonly even more limited. A logical solution is to store large datasets within a relational database and pair this database with an outside application specialized for data visualization, analysis, and automation.

This project specifically is working to facilitate the DCR's (Department of Conservation and Recreation's) data entry, searching, visualization, and analysis process through an R-based application creation tool called Shiny. R is an open source programming language used largely for data statistics and visualization. Shiny is an open source application that allows one to create applications which have a friendly user interface component as well as a server component which uses R to do all the work. A Shiny application can be fully customized by the designer for unique tasks. Features of the DCR Shiny application include easy data searching, water quality time series plotting and analysis, water quality regression analysis, geospatial data visualization and analysis, and more. Through the insight that this application brings, future data collection needs can be better assessed which will direct changes to the current watershed monitoring program.

Automation is a useful, yet underutilized tool that can save an organization much time day to day on decreasing the number of repeated tasks that come along with searching, displaying, and analyzing water quality data. The Application is shared openly through Github and can be ran locally on any computer with two lines of code. Hosting the application online at a URL is also a possibility.

The objective of this project is to create an application that facilitates the data entry, searching, visualization, and analysis of water quality data and other related watershed data.

1.3 Scope of Work (Application Purpose)

The Drinking Water Supply Protection Analysis Application (DWSPA) allows the visualization and analysis of the water quality data and watershed data. There main features of the App include data query and export, data visualization, and data analysis.

Temporal Analysis and spatial analysis (do more (some) reading)

2 Background

2.1 Applications for Watershed Data Science

2.2 Application Developement Frameworks

R - Shiny

Python - Pyramid, Django, Flask)

Other (Ruby on Railsm, node.js)

2.3 Department of Conservation and Recreation Watershed Department

2.3.1 Quabbin and Wachusett Reservoir

The Wachusett Reservoir is located in central Massachusetts and is the main supply of the Boston metropolitan area. The reservoir has a capacity of about 65 billion gallons with a length of 8.4 miles. The reservoir is part of the Massachusetts Water Resource Authority's (MWRA) water supply system which is managed in partnership with the Department of Conservation and Recreation (DCR).

2.3.2 Water Quality Sampling Plan

There are various types of data collected in the watershed. Due to the vast amount of data from various sources the raw data can be in very different formats. Storing data in a central database in similar format is crucial to set up a platform for efficient data analysis. The application provides a connection to this data in the database with an easy-to-use user interface for a human user to interact with the data.

Water quality data includes tributary data grab samples and well as.

Reservoir data includes grab samples, ______, and profile data.

Water quality data related to forestry practices is also collected to compare water quality parameters between managed and unmanaged forests.

This Application will give insight to data collection needs

Data external to the DCR can is also incorporated into the application including hydrological data and meteorological data. Precipitation data is collected from the . River Flow data is collected by the USGS, some gauges in collaboration with the DCR and some independently. Weather data is collected by ______ .

The water quality sampling plan consists of routine sampling at various sites. These sites are Core Sites and EQA sites.

2.3.3 Previous Watershed Studies

3 Application Developement Overview (Method)

3.1 R and RStudio

3.1.1 Shiny

3.1.2 Tidyverse

3.2 Database

3.3 Github

3.4 Developement Strategy/Goals/Practice

Modules and Functions Naming Conventions

3.5 Application Packaging and Deployment

Desktop Shortcut, portable R, portable chrome

4 Application Features and Findings (Results)

4.1 Data Query and Export

It is advantageous, if not essential, to have fast access to water quality and other watershed data. It is beneficial for any scientist or engineer to be able to acess this data with ease. In practice, this is often not the case due to all data not being stored in a location known by all scientists and engineers. Lack of experience with certain technologies can also inhibit a person from being able to access timely data. Although spreadsheets are commonplace among most workplace settings, being able to query this data for the exact data that a perrson is looking for can be timely and troublesome. Queries in a relational database is a better approach to this common task.

The application queries the data based on the user selected input. The user is prompted to first select one or more Site Locations with a map that indicates which sites are selected. The user is then pompted to select the parameters and date range in which data is avaiable for the selected sites. Additional filters can also be applied to the data which include selection of seasons, months, years, flags, storm event, and other eventually meteorlogical and hydrological conditions. Queried data can be exported to a csv file with the click of a button.

4.2 Temporal Analysis

This appliaction allows for water quality visualization of temporal trends and temporal statistical analysis.

4.2.1 Scatter Plot

A scatter plot is avaiable to see a water quality parameters trends over a specified duration of time. The data is queried in a similiar manner as discussed previously and a user is able to plot data from multiple Locations and one or two parameters. The user can chooose to group or facet by Location or Flags. A Facet creates multiple plots, all with similiar mapping techniques, to allow for easy comparison across plots.

The user is given many options to customize how the plot displays the data. the user can choose a log-scale for the Y-axis as well as to start at zero. These options are not applicable for the x-axis becuase the X-axis in date. The user can adjust point size and point color (if the user has not already specified that the group by color). The user can choose form many display themes that are offered in ggplot2. The user can add a horizontal line, vertical line, or floating text anywhere on the plot. The plot automatically creates logical axis labels and a descriptive plot title, which the user can override with custom text or choose to have no labels or title. The user can save a plot with in multiple formats to a specified plot width and plot height. Figures " " are examples of these saved plots. The user can choose to turn on the interactive plot features that are allowed by Plotly which allows the user to hover over a data point for info an also toggle the plot in various ways like zooming in and out.

Temporal Trend lines can be added to the plots to help visualize if there are any temporal trends, if it is not clear when just looking at data points. The trendlines will automatically group in a fashion identical to the points. If the data points on the scatter plots are grouped by Locations as idicated by colors, than the trendlines will also appear grouped by these same colors. Grouping by shape and faceting works similiarly.

Three methods for trendlines are available in this application. The first method is a linear trendline which is the linear line that minimizes the residuals of the fit. The second method is the Loess method which is a . The third method is a Generalized Additive method which . The user can choose to show a confidence ribbon with choices of confidence intervals of 0.90, 0.95, and 0.99. If the user selects a confidence ribbon with 90% confidence, a shaded region will appear on the plot where the data is 95% likely to . This should be used with caution becuase the linear confidence interval does not take into account the seasonal variation, and assumes that all variation is " “.

The user can choose to add a secondary parameter to the plot by introducting a secondary y-axis. This feature allows the user to compare temporal trends of two water quality parameters. A plot comparig of two parameters, on the x-axis and y-axis can usually display a clearer picture of the relationship between two parameters which will be discussed in the Correlation Section. A benefit of the two parameter temporal plot is to keep more of the temporal information and can allow one to visuallly see a more complex trend like a delayed response trend. (Is there a scientific word for this?)

4.2.2 Statistics

Temporal Statistics can be computed with ease which include the maximum, minimum, average, etc statistics and can be grouped by years, months, seasons, parameters, and sites.

Mann-kendall

4.3 Correlation

4.3.1 Scatter Plot

Correlation trend analysis in this application consists of correlation between many types of parameters including water quality, hydrological, and meteorological parameters. Correlation plots can be easily created to see the correlation between parameters with regression lines. correlation statistics can also be computed and shown in an easy to read table. Both of these methods compare parameters taken from the same day and site to minimize temporal and spatial discrepancies.

4.3.2 Pearson Correlation Matrix

Pearson

4.4 Distribution

4.5 Profile Data

4.5.1 Interpolated Color Profile Plot

4.5.2 Profile Line Plot

4.5.3 Profile Statistics

4.6 Phytoplankton

4.7 Geospatial Data Visualization

Spatial trend analysis is incorporated in many locations in the app but primarily lives on the map plot tab. Geospatial plots allows the user to easily compare a parameter statistic across all sites in a visual of plots on their choice of map. Spatial analysis also exists in any tab when multiple sites are chosen for analysis.

4.8 MetaData

4.9 Data Import

The Watershed data Importer Tool (WIT) facilitates importing raw data from multiple sources into the database. Each Data Type has a formatting function script that is written in R to format the data. As more data sources are added or data sources are changed, these can be uploaded into the database. The user is prompted to select dataset type and the user will be shown a list of raw data files in the appropriate dataset type location on their computer. The user then selects a file from their computer and then press a button to format the data. A Warning message will be sent to the user if their was a problem with the data or if the data already exists in the database. After a successful formatting, the user will be able to see the formatted data in a table on the screen and an import button will appear that the user can press to import the data if they are satisfied with how the data looks. Once the data is imported, the raw data file is moved from the unprocessed folder on their computer to the processed folder.

5 Discussion and Reccomendations

5.1 Pros and Cons of Application

Custombility

Upkeep

5.2 Future Work

5.2.1 Meteorological and Hydrological Data

5.2.2 Forestry

5.2.3 Reports

6 Appendix

6.1 WAVE Developer Manual

6.2 WIT Developer Manual

7 References