Yuhang Liang
yuhangl
12-659 Fall 2019 Section A2

Water Quality Correlation Analysis GUI Model

Project Final Report
December 6, 2019

1. Introduction

Water quality analysis and visualization of data from urban landscape lakes is especially necessary so that people can help landscape water keep a healthy state. However, in many cases, water quality control in urban landscape lake tends to focus only on individual indicator. When one indicator in controlled, the other indicator rises and exceed the standard. For example, in the process of pollution control, when the concentration of COD of the lake in controlled, the concentration of total nitrogen may rise during the process. Therefore, a single control of individual indicator will lead to inefficient lake water quality management and rising economic costs. According to the research, since the 1972 U.S. Clean Water Act, government and industry have invested over $ 1trillion to abate water pollution, or $ 100 per person-year, which is a huge cost.[1] This reminds us of the importance of water quality analysis. What's more, interactivity is another important point of this project. It is much better for the analysis methods to be packaged into finished software with GUI, which is lacked now, so that people who lack the programming foundation can also complete the image output just by inputting data.

By understanding the changing characteristics of the water environment, we can better help us to control pollution. In the process of water quality analysis, we need to find the correlation between different chemical indicators in the water, and the correlation between water environmental factors and pathogen indicators. These related properties allow us to understand the changes in water quality and to dynamically manage the landscape lakes in the city. In addition, software of this analysis methods can help more people use the power of math easier.

2. Background

This report concentrates on two points, correlation analysis and graphics user interface. For correlation analysis, this project uses statistical procedure, principle component analysis (PCA) and redundancy analysis (RDA), to analyze data. For graphics user interface, the project uses App Designer as the environment for building apps.

2.1 Detailed background about correlation analysis

In our lives, we often encounter relatively large databases, some of which might be correlated. Because of the size of the data, it is not easy for us to find the correlation between different variables. Thanks to the redundancy, brought by the correlation, in the information that can be by the data set, we can do correlation analysis on the data. In order to reduce the computational and cost complexities, we use PCA to transform the original variables to the linear combination of these variables which are independent. The reason why we use PCA in this project is that PCA is the simplest of the eigenvector-based multivariate analyses and it is closely related to factor analysis.[2] The correlation analysis of the database we use in the project is a typical factor analysis. We can use PCA to explore the correlation between different chemical factors.

In terms of RDA, it is actually an extension of the iterative algorithm of PCA. It allows studying the relationship between two tables of variables Y and X. In this project, the variables represent water environmental factors and water pathogen factors separately. And their relation can be clearly by the RDA algorithm.

Based on the obtained correlation data, the project looks for the linear relationship of some closely related indicators, so that the project's products are more comprehensive in function.

2.2 Detailed background about graphics user interface

The GUI has now played an increasingly important role. It can improve the efficiency of people using mathematical tools. This project is designing a well-interactive GUI that allows users to import the specified data to get the desired statistical analysis results. Therefore, the project is now using App Designer as the environment for building apps in Matlab. This environment allows to quickly move between visual design in the canvas and code development in an integrated version of the MATLAB Editor, which is suitable for the project.

3. Implementation

The objective of the project is to create a complete app with statistical capabilities. In terms of data analysis, as long as the data is input, the app can output the relevant principal component analysis (PCA), redundancy analysis (RDA) and other results, and draw related images. The data is not limited to the lake water quality data used by the project, but can also use data other than CEE fields. On the GUI side, an input interface that can adjust the size of the input data will be completed, which separates environmental factors and biological factors when doing RDA. This interface can select output forms such as image output and digital output.

3.1 Solution Methods

The completion of this project mainly requires two parts, the design of the algorithm and the acquisition of sample data. The main algorithms that used in this project is PCA and RDA. The data on which the project is based is from three landscape lakes in different cities in China.

3.1.1 Algorithm

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables (entities each of which takes on various numerical values) into a set of values of linearly uncorrelated variables called principal components.[2]

Let Y_{ij} be the value of variable j (j=1,…,N) for observation i (i=1,…,M). Most ordination techniques create linear combinations of the variables:

$$Z_{i1} = c_{11}Y_{i1} + c_{12}Y_{i2} + \cdots + c_{1N}Y_{iN}$$

Using an iterative algorithm, we can explain PCA as the following steps: 1. Normalize the variables in Y. 2. Obtain initial scores z. 3. Calculate new loadings: $c = Y'z'$. 4. Calculate new scores: $z = Yc$. 5. For second and higher axes: Make z uncorrelated with previous axes using a regression analysis. 6. Scale z to unit variance: $z^* = z/\lambda$, where λ is the standard deviation of the scores. Set z equal to z^*. 7. Repeat steps 2 to 6 until convergence. After convergence, divide λ by $M - 1$. [5]

Redundancy analysis (RDA) is a method to extract and summarize the variation in a set of response variables that can be explained by a set of explanatory variables. More accurately, RDA is a direct gradient analysis technique which summarizes linear relationships between components of response variables that are "redundant" with (i.e. "explained" by) a set of explanatory variables.[4]

PCA calculates the first principal component as

$$Z_{i1} = c_{11}Y_{i1} + c_{12}Y_{i2} + \cdots + c_{1N}Y_{iN}$$

Redundancy analysis is a sort of PCA that the components are linear functions of the explanatory variables:

$$Z_{i1} = a_{11}X_{i1} + a_{12}X_{i2} + \cdots + c_{1q}X_{iQ}$$

Hence, the axes in RDA are not only a linear combination of the response variables, but also of the explanatory variables. Further axes are obtained in the same way.[5] This technique requires an explicit division of the variable into response and explanatory variables. In this project, we consider water environmental factors as explanatory variables and water pathogen factors as response variables.

3.1.2 Data

The database contains monitoring data for twelve consecutive months for three lakes. It includes 22 water environmental factors and 8 water pathogen factors. Water environmental factors include basic physical indicators (temperature, visibility, etc.) and chemical indicators (total nitrogen concentration, total phosphorus concentration, COD concentration, etc.). The type of the database is csv file, which is used for analysis by PCA, RDA, etc.

3.2 Program Design

3.2.1 Code

The code of this project is mainly divided into two major blocks: the data analysis part and the app design part. Data analysis included data import section, trend analysis, PCA analysis, and RDA analysis. The app design mainly includes the design of the user interface, the connection of various operation options on the interface with data analysis functions. The diagram that shows the overview of this project is shown in Figure 1.

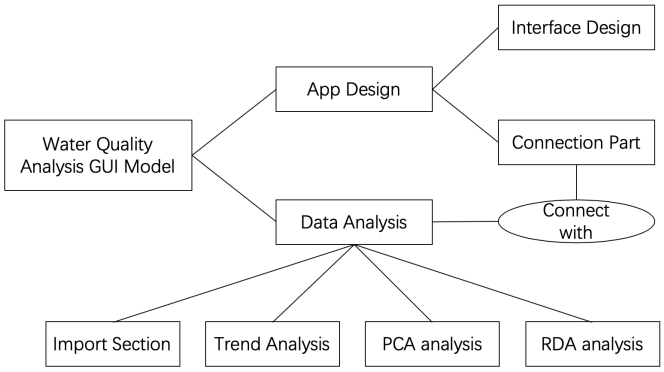

Figure 1: Overview of the Project

3.2.2 Toolbox

The toolbox that this project mainly uses is Fathom Toolbox for Matlab, which is a collection of statistical functions written by USF's College of Marine Science. This toolbox supports Redundancy Analysis (RDA) and Principal Components Analysis (PCA), which are related to this project.

3.2.3 Combination

Regarding the combination of the toolbox and the project body code, in this project, the Fathom Toolbox used is a function with data analysis capability. The project used parts of the code of the PCA and RDA functions in the Fathom Toolbox and modified them to fit the needs of the project. And ultimately the functions required for the project are formed and invoked in the GUI interface.

3.3 Logic Progress

The ultimate goal of the project is to create a GUI model with water quality analysis capabilities. The way to achieve this is to properly write each data analysis function into a GUI. Therefore, the project firstly writes each function needed, and creates an interface on the GUI according to the requirements of each function, and finally imports each function into GUI. Therefore, there are two major junks of code in this project, data analysis and app design. The diagram that shows how these junks flow together is shown in Figure 1.

3.3.1 Data Analysis

Data analysis included data import section, trend analysis, PCA analysis, and RDA analysis. For data import section, its function is mainly to import the database with the file format, and then store the corresponding data in different variables in Matlab, providing preconditions for the functions to call the variable. For trend analysis, its function is to plot the time-series data. For PCA analysis, its function is to perform PCA analysis on the imported database and draw corresponding images. For RDA analysis, its function is to perform RDA analysis on the imported database and draw the corresponding image.

3.3.2 App Design

The code in the App Design section is mainly divided into two parts, the interface design and the connection of objects on interface and different functions. For the design of interface, mainly through the

App Design Environment, this part of the code is automatically generated by Matlab. For the second point, in this project, the main implementation is to import the data into the app, and bind the call command of each function to the corresponding object on the interface.

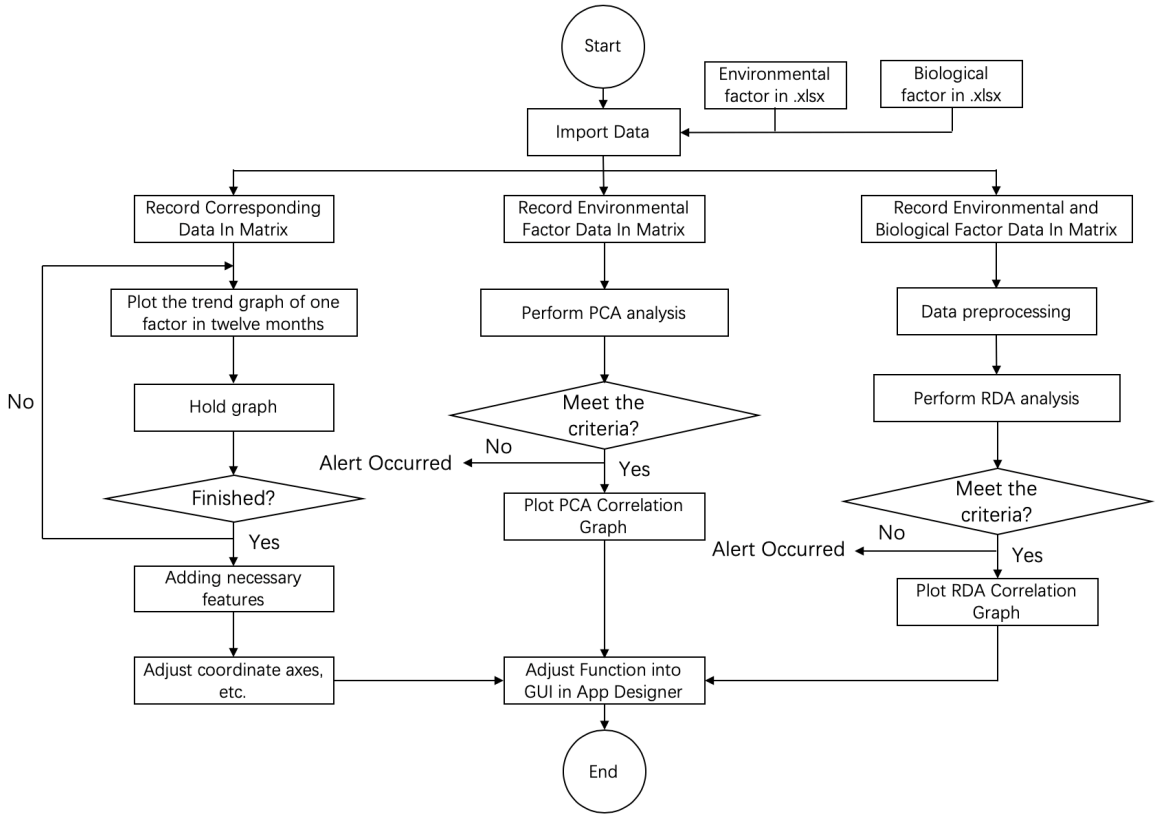

Figure 2: Flow Chart

4. Conclusion

4.1 Project Production

The project completed a water quality analysis model in accordance with the goals of the project. Figure.3 is the user interface of the model. This interface has a friendly user experience. Users can load the environmental factor database and biological factor database separately through the load buttons. The model will show the current environment of the data in the dropdown box. By clicking the buttons of different analysis methods, the model performs corresponding analysis according to the environment selected by the user. Through the color switch of the button, the model can clearly show what analysis is currently being performed.

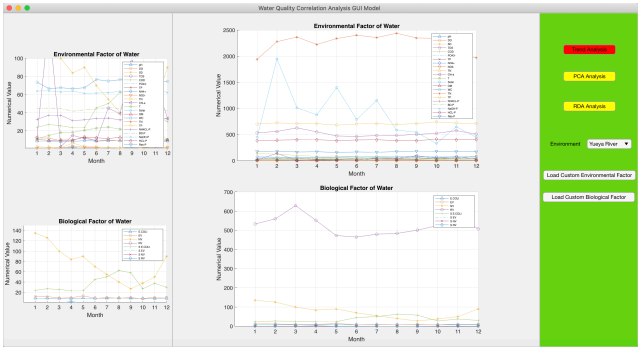

Figure 3: User Interface of Model

In the Figure.3, the trend analysis is currently performed. The middle part of the interface shows a trend chart of all data, and the left part shows an enlarged view of the middle part of the data to achieve the purpose of clearly expressing the data trend.

In the PCA part, the model can implement principle component analysis in different environments. Figure 4 is the PCA analysis results for the default sample. In the Figure 4, the first main component accounts for 33.11% of the contribution rate, and the second main component accounts for 17.15% of the contribution rate.

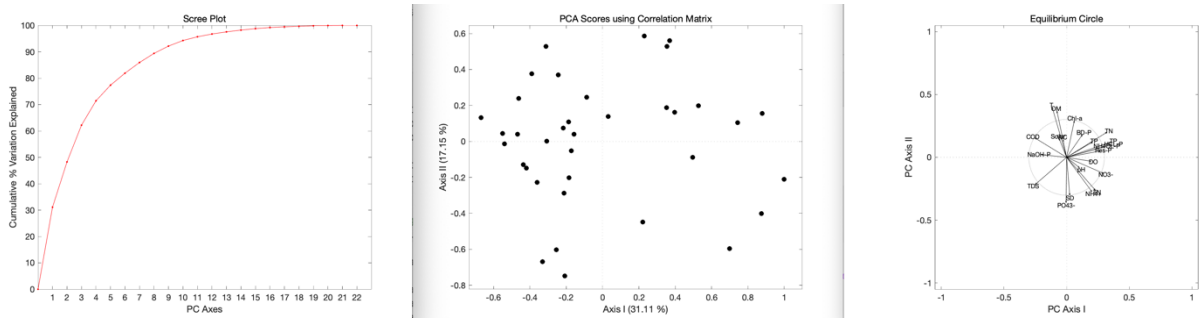

Figure 4: Result of PCA

In the RDA part, the current model can analyze the numerical value of redundancy analysis in different environments, and can read the corresponding RDA result graph.

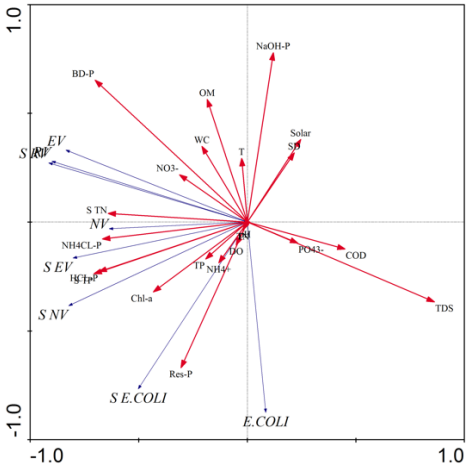

Figure 5: Result of RDA

4.2 Project Gains

In Matlab programming, based on the knowledge covered in lectures, I am able to use Matlab skills, including flow control, package import and use. At the same time, I developed good programming habits, such as reasonable indentation, comment and classification and calling of functions. In addition, I learn to use the Matlab programming language to achieve mathematical principles, and further data analysis capabilities. Finally, at present I can use Matlab to create a GUI interface and visualize the model so that the app can perform the expected data analysis functions. And I have a deeper understanding of OOP programming.

In terms of algorithms, I studyed the mathematical principles of both PCA and RDA algorithms, analyze the results of PCA and RDA, and used these two methods to analyze the correlation of different types of data.

4.3 Future Work

If I have more time on this project, I will implement the RDA function of the model more beautifully and perfectly. In addition, I will add the function of water quality prediction, which will make my project more practical.

Reference

1. David A. Keiser and Joseph S. Shapiro. *Consequences of The Clean Water Act and the Demand for Water Quality*. Yale University, January 2017.
2. https://www.quora.com/When-and-where-do-we-use-PCA
3. https://en.wikipedia.org/wiki/Principal_component_analysis#Details.
4. https://snoyadihnif.wordpress.com/2013/06/10/flow-chart/.
5. Alain F. Zuur, Elena N.leno and Graham M.Smith. *Analysing Ecological Data*.