7-1-2022

Open Innovation Web-Based Platform for Evaluation of Water Quality Based on Big Data Analysis

Xiaofang Han
Huazhong Agricultural University

Hong Shen
Institute of Hydrobiology, Chinese Academy of Sciences

Hongqing Hu
Huazhong Agricultural University

Jerry Gao
San Jose State University, jerry.gao@sjsu.edu

Recommended Citation

Xiaofang Han, Hong Shen, Hongqing Hu, and Jerry Gao. "Open Innovation Web-Based Platform for Evaluation of Water Quality Based on Big Data Analysis" *Sustainability (Switzerland)* (2022). https://doi.org/10.3390/su14148811

sustainability

Article

Open Innovation Web-Based Platform for Evaluation of Water Quality Based on Big Data Analysis

Xiaofang Han [1,2], Hong Shen [3], Hongqing Hu [1,*] and Jerry Gao [4,*]

[1] College of Resources and Environment, Huazhong Agricultural University, Wuhan 430070, China; xfhan@jhun.edu.cn
[2] School of Environment and Health, Jianghan University, Wuhan 430056, China
[3] Donghu Experimental Station of Lake Ecosystems, Cern Sub-Center of Aquatic Ecosystems, Institute of Hydrobiology, Chinese Academy of Sciences, Wuhan 430072, China; hongshen@ihb.ac.cn
[4] Computer Engineering Department, San Jose State University, San Jose, CA 95192, USA
[*] Correspondence: hqhu@mail.hzau.edu.cn (H.H.); jerry.gao@sjsu.edu (J.G.)

Abstract: There are many models presented that assess water quality. However, the applications of the models are limited due to the difficulty of preparing input data and interpreting model output. In this paper, we developed a Web-based platform to assist researchers in analyzing water quality. The data from sensors can be automatically imported to the platform according to the configured information of data structures. The platform also provides conventional methods and big data methods for the users to analyze water quality. Moreover, the users can choose the water quality parameters according to the water usage. The presented platform can show the model output in a text format and a graphic format, which allows for the analysis to be better understood by the user. The platform integrates the input, analysis, and output together well and brings great convenience to the research on water quality.

Keywords: water quality; big data analysis; application; platform

check for updates

Citation: Han, X.; Shen, H.; Hu, H.; Gao, J. Open Innovation Web-Based Platform for Evaluation of Water Quality Based on Big Data Analysis. *Sustainability* **2022**, *14*, 8811. https://doi.org/10.3390/su14148811

Academic Editor: Ozgur Kisi

Received: 21 May 2022
Accepted: 5 July 2022
Published: 19 July 2022

1. Introduction

Human activities, such as urbanization, agriculture, etc., often influence water quality. The water sources have been degraded by human and enterprise waste so much that there is a scarcity of fresh water. Water quality refers to the chemical, physical, biological, and radiological characteristics of water [1]. The requirements of water quality are related to the water usage. Thus, many standards on water quality assessment have been published. The most common standards used to assess water quality relate to health of ecosystems, safety of human contact, and drinking water. Although there are many studies related to water quality using conventional methods, there is no proper analysis of water polluting agents and its condition. The main shortcoming of conventional methods is that the methods usually focus on one or several special points on water quality and hardly provide a comprehensive analysis result.

Due to recent technological advancements, especially in big data technology, many new water-quality assessment methods based on big data have been proposed for water management and quality prediction. Researchers have recognized the necessity of integrating water quality models with the best available technology to overcome the barriers that hinder the use of many models. One problem is that the applications of some scientific models are limited due to the difficulty of preparing input data and interpreting model output. In this paper, we developed a Web-based platform to assist researchers in analyzing water quality that automatically imports the input data according to the user configuration. The developed platform uses various suitable technologies that can perform detailed analytics and predictions on water data. We developed a big data engine to efficiently analyze the water quality data. Aided by the platform, environmentalists can find a better solution to improve water quality.

The rest of the paper is organized as follows. In Section 2, the background and related work on water evaluation are reviewed. The system analysis and design on the proposed platform for water evaluation is described in Section 3. Section 4 discusses the analytics module in the platform. In Section 5, the application of the platform is reported. Finally, our conclusions are drawn in Section 6.

2. Background and Related Work

2.1. Background

The most common standards used to assess water quality relate to health of ecosystems, safety of human contact, and drinking water. Table 1 shows the main water quality standards and their usages.

Table 1. Water quality standard and their usages.

Water Quality Index	Water Usage
Brown (NSF-*WQI*)	General
Nemerow and sumitomos index (US Environmental Protection Agency)	Human, indirect, remote contact use
Prati's index	General
DEininger and landwehr's	Public Water Supply (PWS)
McDuffie and Haney's	River pollution index
Dinius' water quality index	Impact of pollution control efforts
O'Connor's Indices	Fish and Wildlife (FAWL) and PWS
Walski And Parker's index	recreation water (swimming and fishing)
Stoner's index	PWS and irrigation water

Although different water quality standards select different parameters to evaluate water quality based on water usage, the following parameters are generally used for analyzing the water quality [2].

The water quality index includes chlorophyll A, dissolved oxygen, pH, salinity, turbidity, water temperature, nitrates, phosphates, water level and chlorine. (See supporting information).

The water quality assessment is based on a four-step process: (i) Select water quality evaluation criteria; (ii) Assemble existing water quality data; (iii) Evaluate water quality data for selected parameters; (iv) Draw conclusions. According to the above conventional method, it is easy to determine the current status of water quality. However, we only depend on the traditional methods; it is difficult for us to predict the future condition of water quality and find the potential cause of water quality decline.

2.2. Related Works

A lot of research on water quality has been done in the past years. The methods of analyzing water quality can be divided into two categories: the conventional methods and the big data methods. As we know, different water quality standards should be used to evaluate the water according to its usage. To detect and analyze gradual and abrupt trends in water quality in Huai River, Chemical Oxygen Demand (COD) and ammonia-nitrogen (NH3-N) are selected as the metrics [3]. Mann–Kendall, Theil–Sen and Pettitt tests are used to process the data from the monitor station. In the evaluation of water quality of the Amazon area in Brazil, the statistical technique using principal component analysis is used [4]. The water quality index used here includes chloride, temperature, electrical conductivity, pH, Total Suspended Solids (TSS), ammonia (NH4), nitrate (NO3), chlorophyll, turbidity (Turb.), dissolved oxygen, phosphates (PO4), and upstream water level. Since so much research on water quality evaluation has been proposed, we selected some prototype evaluation models or methods for drinking water, industrial use water and coastal water, etc. The models and the water quality parameters used by them are listed in Table 2.

Table 2. Some conventional models and methods for water quality evaluation.

Paper ID	Purposes	Model/Methods	Parameters/Metrics	Data Source
[5]	Evaluate coastal waters for ecological and recreational value	EEA 2001	NH^{4+}, NO^{2-} and NO^{3-}	Historical data
[6]	Water deterioration in southern Portugal	Evaporation and salinity model	temperature and salinity	Data stations
[7]	Seawater desalination capacities	Environmental Impact Assessment (EIA)	Temperature, Density and Salinity	World Health Organization
[8]	Predict ocean water quality in Hilo Bay	Artificial neural network	Chlorophyll, Dissolved Oxygen, Salinity, Turbidity and Temperature	Data station in Hilo Bay
[9]	Analyze water quality restoration in Guanabara Bay	Eutrophication models	Chlorophyll, phosphorous, nitrogen, BOD, DC, and PC	Japan International Cooperation Agency
[10]	Eutrophication and degradation of coastal waters	WASP Model	Salinity, Dissolved Oxygen, Nitrogen, Ammonia, Organic Nitrogen, Phosphorus	
[11]	Analyze water quality of Chesapeake Bay	Hydrodynamic model	Dissolved Oxygen, Nitrite, Ammonia, Chlorophyll, Total phosphate	Chesapeake Bay
[12]	Evaluate pollution effects on river water quality	Sampling and Cluster	Iron, Flow rate, hardness, bicarbonate, potassium, magnesium	Water sample

In recent years, the big data method has been widely applied to many domains. Some researchers also use big data methods to analyze water quality. Table 3 shows some topics on water quality using big data methods.

Table 3. Some big data methods for water quality evaluation.

Paper ID	Purpose	Algorithms
[13]	Determine the environmental quality of a drinking water reservoir by remote sensing	multiple Regression Analysis
[14]	Evaluate water quality trends in an estuary	Weighted regression
[15]	Evaluate groundwater quality	Multiple linear regression
[16]	Water quality classification	Machine leaning algorithms
[17]	Evaluate water quality of canals in Bangkok	K-means
[18]	Heavy metal pollution assessment in the Shur River	Support Vector Machine
[19]	Water quality forecasting in an agricultural nonpoint source polluted river	Support Vector Machine
[20]	Prediction of BOD5 in river systems	Support Vector Machine

According to the literary survey on the water quality problem, it is found that K-means, SVM and multiple linear regression (MLR) are the three most-used big data approaches to analyze water quality.

In our preliminary research, we found out that most of the water quality research conducted depends on the traditional or laboratory approach in which a sample of water is tested against one or more tests to evaluate the quality of water. These approaches involve manual collection and assessment of raw data. M. A. Tirabassi [2] formulated a mathematical model using mathematical statistics to predict river water quality without using the chemical, biological, and physical relationships. His work focuses on a "black box" approach where a known input can be used to accurately and reliably predict the output. L. Hu, et al. [3], has talked about Grey Relational Analysis (GRA), which is based on the distance of a point to the interval. It is a simple method that has been used for the assessment of drinking water quality in Jiaozuo River, China. However, these traditional methods are confusing in nature and have several shortcomings such as:

1. The existence of complex mathematical calculations;
2. Equal treatment to the old data and new data;
3. Difficult prediction due to overlap of multiple variables.

To overcome these problems, we developed a Web-based platform to perform analytics on various water data to provide useful and meaningful information. The platform uses various suitable technologies that can perform detailed analytics and predictions on water data. A big data engine is also provided in our platform to efficiently analyze the water quality data. Thus, our platform is a good tool for researchers to find a solution to improve water quality.

3. System Analysis and Design

3.1. System Overview

In this paper, we will develop a Web-based application to analyze water quality. The proposed application combines the traditional method and big data method to evaluate the water quality. The main functions of the system are described as follows:

- Data Collection: The application can automatically collect data from the configured data source. A configured panel is provided to the system manager to set how to collect data.
- Data Extraction: The application can extract the corresponding data according to the purpose of analysis. For example, it can extract the data by region, year, and water quality parameters, etc.
- Data Analysis: The application uses different methods to analyze data. Some traditional methods and big data methods are implemented here, which are provided to the users. Moreover, data analysis can be done from different perspectives, such as water region, time period, and overall water data.
- Data Visualization: The analysis results should be shown in a friendly way to the users. The data visualization module provides an easy way to show the results and help users better understand the results. For example, if there are some certain points in the area that lead to heavy pollution, they are shown in the map by highlights or a special color.
- Comparing Different Models: In this application, different models or algorithms are used to evaluate the water quality. Different models would give different analysis results with different accuracy; hence, the application will provide a comparison of these different models.
- Prediction: This is an additional feature of the application. The application provides prediction results on the tendency or future result based on the current water quality.
- Finding Causes: The application can help the users to find the cause of the anomalies in the data and the source of pollution.

3.2. System Infrastructure

The application system is designed based on a three-tier structure, i.e., client, database, and middle tier. The three-tier architecture is shown in Figure 1.

The client tier consists of the following modules: data collection module, data analysis module, data visualization and service quality module. In the data collection module, users can collect the data from the website by choosing the durations, type of parameters, region, and data sources. Then, the collected data are cleaned and validated. The data analysis module provides two types of approaches for performing data analysis: conventional methods and big data methods. The data visualization module provides maps and different graphs, such as bar graphs and line charts, to visualize the data and analysis results. The service quality module is used to keep track of the product quality. It also stores the user activity records and provides some functions such as data service management and security.

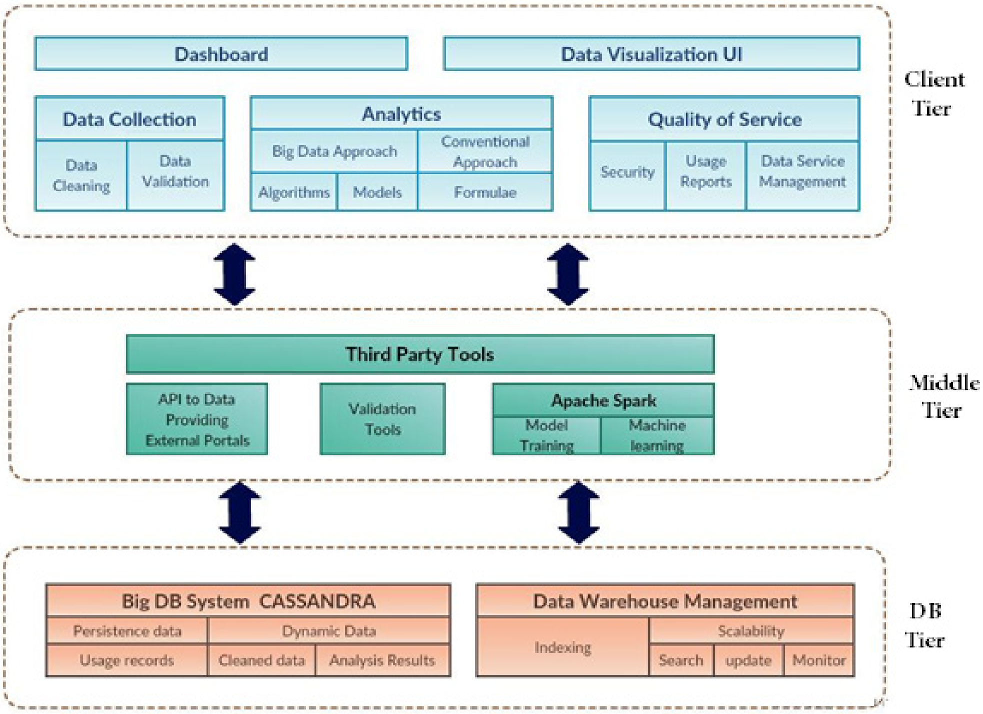


Figure 1. System architecture of the application.

The middle tier is the connection between the client tier and the DB tier. The middle tier provides the logical processing for the data. The middle tier mainly consists of the third-party tools used by the applications, such as the third-party API to external portals, validation tools and Apache Spark. The middle tier can be seen by a third-party tool container. By making full use of the third-party tool, the application obtains many basic functions in data collection, validation, visualization, and analyzing, which provides support effectively to the functions in the client tier.

The DB tier is in the bottom of the system architecture, which is the foundation of the application. The DB tier consists of the database and its management. There are two databases in our application: one is used to store the user analytics data, and the other is the scalable database to keep sensor data. The data warehouse management module is an important part to making the database efficient by using indexing, reference key and search key, etc. It is also responsible for scalability and performance of the warehouse.

3.3. Class Diagram

According to the three-tier structure, we find and define the classes in the three tiers. Then, we detailed the interaction between these classes. Figure 2 shows the class diagram and the interaction between them. The description on the class diagram is as follows:

- User: User class keeps track of all the users' login and logout details. A user can extract data and get the analysis results;
- Collect Sensor Data: User extracts data using this class. Desired request of data is first looked up in the big data database Cassandra using search and reference keys. If not found, third-party APIs are used to extract the data from the data sources. Data is then cleansed using validation tools and stored in the database for further use;
- Conventional Analysis: User can perform conventional analysis using this class. This provides various formulae to perform the analysis;
- Big Data Analysis: To see the results of big data analysis, this class is invoked. This provides different models such as the analysis model and predictive model, using a variety of algorithms to perform analysis;

- Data Visualization: This class provides various libraries to draw different charts on the basis of what chart type the user wants to use for analysis;
- Spark: Big Data Analysis class uses the third-party tool, Spark. Any functionality and programming associated with the tool Spark are maintained and written in class "Spark;"
- Admin: Admin class is made for admin-specific functionalities. An admin can add different models, diagnose the applications, modify functionalities related to database and see the user analytics reports;
- Usage Report: User analytics reports are maintained in this class. This class uses the service database made for the application and stores the regular user activities. This class is mainly accessed by admin to track the activities of the application and user;
- Validation Record: When data are collected, the raw data are cleansed, and only clean data are kept in the database. The records of the cleaning are kept using this class. It records the results of how much data was missing and truncated from the raw data. This class is useful to find the quality of the data source. For example, for a particular database, 99% of the data is always truncated due to garbage values or missing data, then the admin could decide whether to use that data source in the future or not;
- Data Service Management: This class provides database-specific activities. All the database operations such as select, update, insert and delete are handled by this class. The application consists of two databases. One is for storing data and the other is for analytics. Both of these databases are maintained in this class.

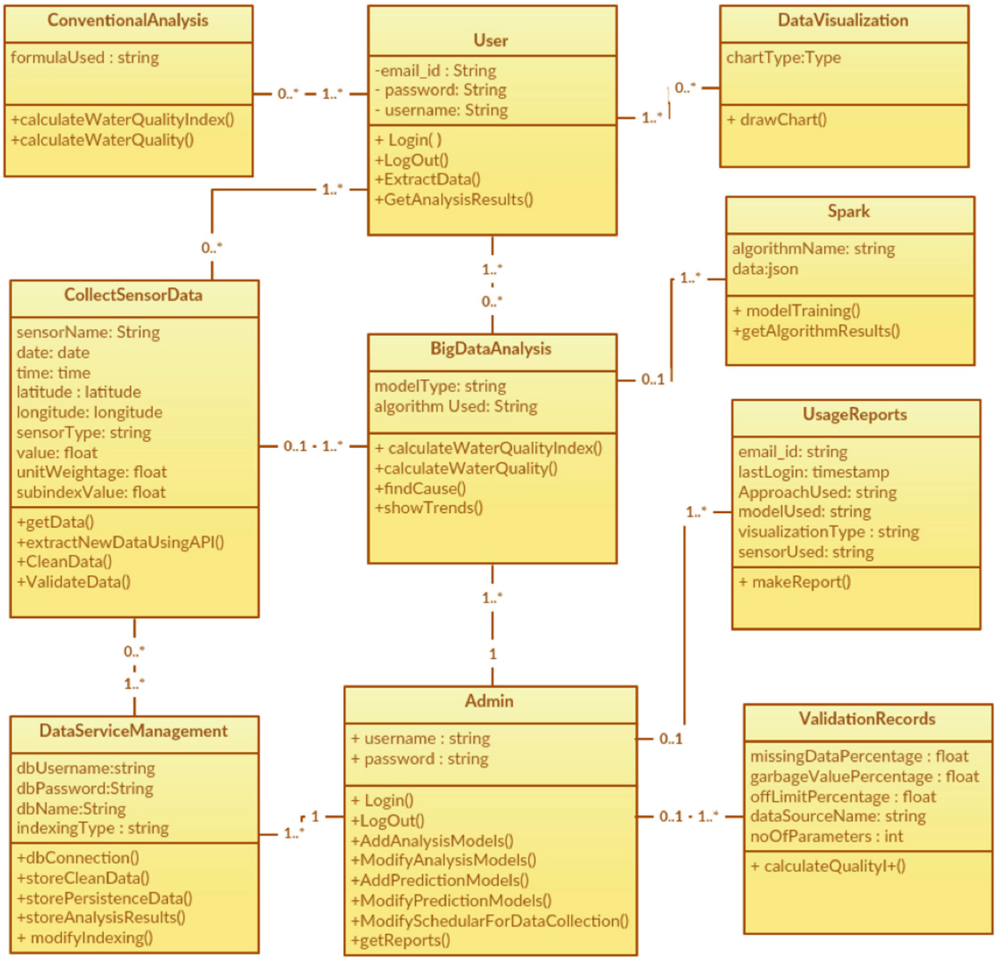


Figure 2. Class diagram. "*"means that it must appear at least once.

3.4. System Data and Database Design

In our application, the relational database and big data database are both used for different purposes. The service data, which mainly include the user information and their querying records, are stored in the relational database. The water quality data, which are collected from the sensors, are stored in the big data database.

3.4.1. Big Data Warehouse Database Design

In our application, all data from the sensors include sensor data value, sensor data type, sensor location, and time. The main process of data flow is as follows: the sensor data are automatically collected according to the system manager's configuration. Then, the collected data are cleaned and stored in the data warehouse. When users require specific data, the application will try to find the data from database and then output the data. If the data cannot be found, the system will call the collection API to obtain the data again. In addition, the data warehouse will also store the data analysis results. The indexed B-tree technique is used when storing the data.

Since the data collected from sensors are the non-structure data, we chose Cassandra database as our big data database warehouse. Figure 3 shows the conceptual data model diagram for Cassandra design. This design is originally from the big data database ER diagram. Then, according to this conceptual data model and mapping rules, the infrastructure for Cassandra is obtained and shown in Figure 4.

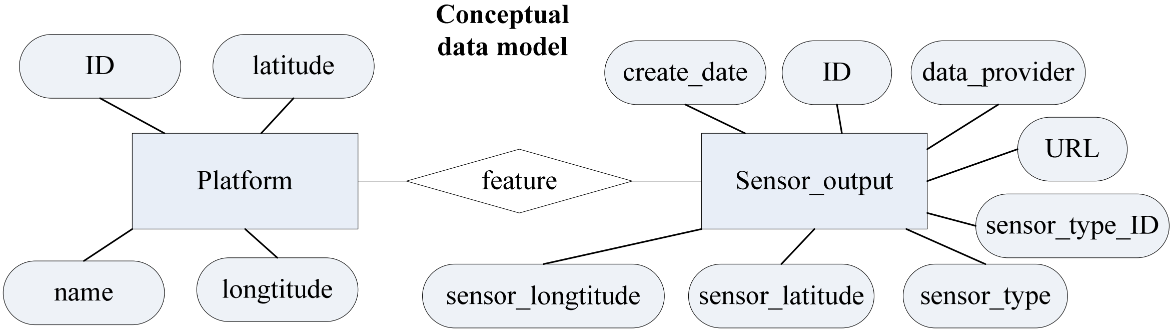
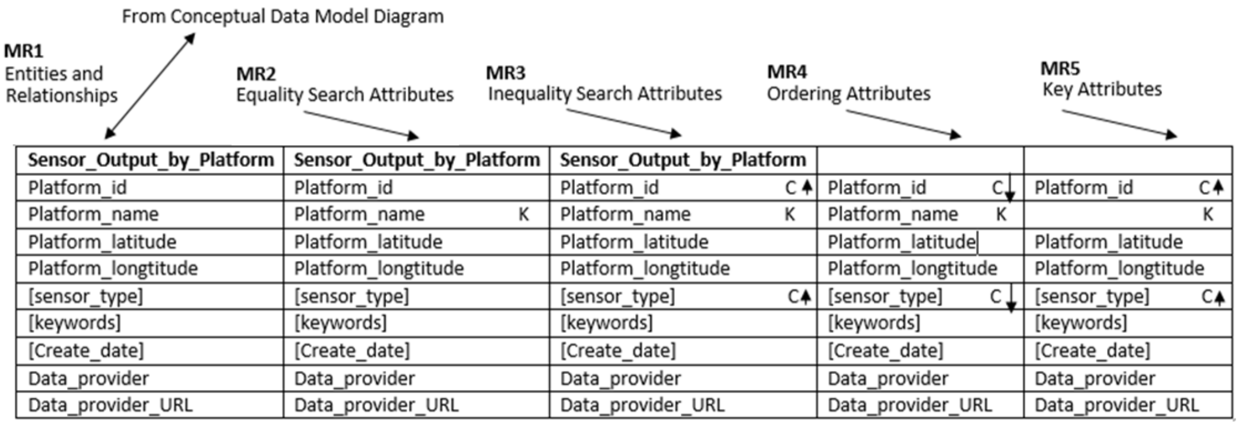


Figure 3. Conceptual data model diagram for Cassandra.


From Conceptual Data Model Diagram
MR1
Entities and Relationships
MR2
Equality Search Attributes
MR3
Inequality Search Attributes
MR4
Ordering Attributes
MR5
Key Attributes

Sensor_Output_by_Platform	Sensor_Output_by_Platform	Sensor_Output_by_Platform		
Platform_id	Platform_id	Platform_id C↑	Platform_id C↓	Platform_id C↑
Platform_name	Platform_name K	Platform_name K	Platform_name K	K
Platform_latitude	Platform_latitude	Platform_latitude	Platform_latitude	Platform_latitude
Platform_longtitude	Platform_longtitude	Platform_longtitude	Platform_longtitude	Platform_longtitude
[sensor_type]	[sensor_type]	[sensor_type] C↑	[sensor_type] C↓	[sensor_type] C↑
[keywords]	[keywords]	[keywords]	[keywords]	[keywords]
[Create_date]	[Create_date]	[Create_date]	[Create_date]	[Create_date]
Data_provider	Data_provider	Data_provider	Data_provider	Data_provider
Data_provider_URL	Data_provider_URL	Data_provider_URL	Data_provider_URL	Data_provider_URL

Figure 4. The Cassandra infrastructure for the application.

3.4.2. Relation Database Design

The relational database in our system is used to store the service data. The service data include the user information and their querying records. Once a user has created a login name and password from the service website, the service database will be automatically assigned a user ID. Then, the users can create their project for evaluating water quality, which will automatically be assigned a project ID. The project table includes what users are querying for, such as duration, location, platform, and its sensor, etc. In addition, one user's project can have more than one sensor output query result as the user can query different parameters. The detail result will be in the sensor_record table, which has been cleaned and analyzed before putting it in the database. Moreover, an administrator can access the usage report including who was using the service, what water quality parameter has been queried, what the duration is, etc.

According to the above analysis, we designed the infrastructure of the relational database for the service data. Figure 5 shows the database design for the service data.

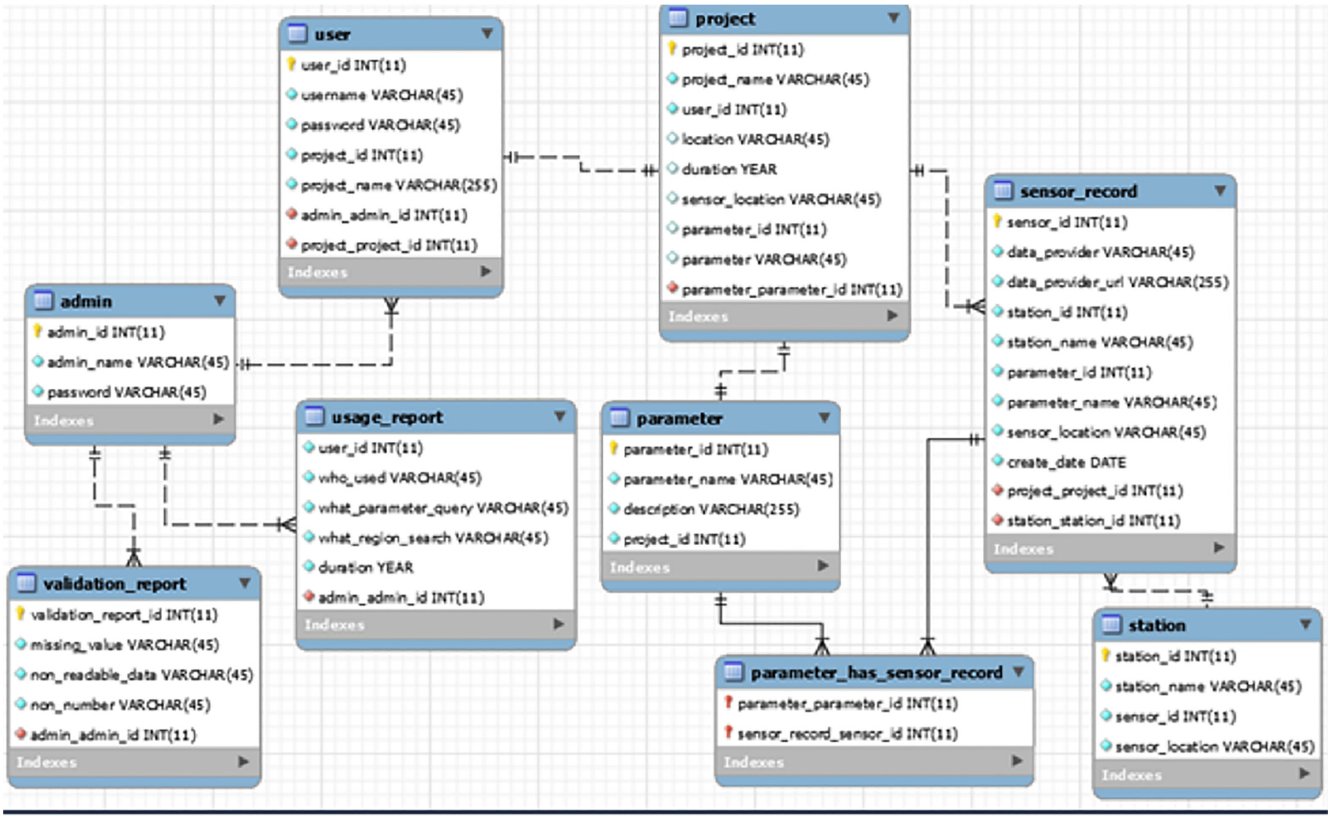


Figure 5. Database design for the service data.

4. Analytics Module—Water Quality Assess Modeling

The analytics module is an important part of this system where overall water quality is calculated and analyzed. The analytics module consists of some analytics models. The analysis methods used here can be divided into two categories, i.e., conventional approaches and big data approaches.

4.1. Conventional Model

In the conventional approaches, NSF-*WQI* is the most-used method for evaluating water quality. It was developed by the National Sanitation Foundation (NSF) in 1970. A total of nine water quality parameters such as temperature, pH, turbidity, fecal coliform, dissolved oxygen, biochemical oxygen demand, total phosphates, nitrates, and total solids are used to evaluate water quality. Collected water quality data are transferred to a weighting curve chart, where a numerical value of Q_i is obtained. The detailed description on NSF-*WQI* can be found in [20]. The mathematical expression for NSF-*WQI* is as follows:

$$WQI = \sum_{i=1}^{n} Q_i W_i \tag{1}$$

where Q_i is the ith number of water quality parameter; W_i is the weight associated with ith water quality parameter, n is the total number of parameters.

However, NSF-*WQI* only represents the general water quality and does not represent specific use of the water. Moreover, it does not provide the method of how to process the lack of data or missing data. To remedy the above shortcomings, we used another water evaluation index from the Canadian Council of Ministers of the Environment (CCME) in our application. In the water evaluation of CCME, the major uses of water include drinking, recreation, agriculture, aquatic life, and industrial use. Following expressions F_1, F_2, and F_3 are used to determine the CCME *WQI* for our study [21].

$$F_1 = \frac{\text{number of failed variables}}{\text{total number of variables}} \times 100 \tag{2}$$

$$F_2 = \frac{\text{number of failed tests}}{\text{total number of tests}} \times 100 \tag{3}$$

$$F_3 = \frac{nse}{0.01nse+0.01} \tag{4}$$

Here, F_1 (scope) represents the percentage of variables that do not meet their objective at least once during the time period under consideration ("failed variables"). F_2 (frequency) represents the percentage of individual tests that do not meet objective (failed tests). F_3 (amplitude) represents the amount by which failed test values do not meet their objectives. F_3 is calculated by an asymptotic function that scales the normalized sum of *excursions* from objectives (*nse*) to yield a range between 0 and 100. *nse* is calculated as below:

$$nse = \frac{\sum_{i=1}^{n} excursion\ i}{\#tests} \tag{5}$$

The number of times by which an individual concentration is greater than (or less than, when the objective is a minimum) the objective is termed an "*excursion*" and expressed as follows.

$$excursion\ i = \frac{\text{objective } j}{(\text{Failed test value } i)^{-1}} \tag{6}$$

F_3 is then calculated by an asymptotic function that scales the normalized sum of *excursions* from objectives (*nse*) to yield a range between 0 and 100.

Finally, the CCME *WQI* is calculated as below:

$$WQI = 100 - \frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \tag{7}$$

According to the value of *WQI*, the water quality is shown in Table 4.

Table 4. CCMI *WQI*.

CCME Index Value	Water Quality
95–100	Excellent
80–94	Good
60–79	Fair
45–59	Marginal
0–44	Poor

The *WQI* method is widely used in surface water and groundwater assessment. According to Ahmed M [22], 48 groundwater samples were collected and analyzed for 12 physicochemical parameters. Multivariate statistical approaches, including a correlation matrix, factor analysis, and hierarchical cluster analysis, were applied to differentiate the source of the water quality variation and determine the cause of groundwater deterioration. The *WQI* was applied according to the chemical drinking-water quality standards of the World Health Organization (WHO) with respect to the 12 parameters measured to evaluate the suitability of the groundwater for human consumption.

4.2. Big Data Modeling and Analysis

Water quality analysis is a classification/clustering problem since we need to find the regions and classify them to find which comes under the ideal recommended conditions, above/below the recommended values, worst conditions regions, etc., on the basis of parameters and time range selected. In the proposed system, some machine learning algorithms are used for the analysis of water quality. Our application provides some frequently used algorithms in big data analysis, i.e., K-means, Support Vector Machine and Multiple Linear Regression.

4.2.1. K-Means Clustering

After comparisons, we found that K-means clustering algorithm is the best for the water quality classification problem. The K-means algorithm is a cluster analysis algorithm used as a classification or partitioning algorithm. The K-means algorithm defines a random cluster centroid at first according to the initial parameters and criteria selected. Every observation is then added to the cluster according to the proximity to the cluster centroid. The clusters are then re-analyzed to determine the new centroid point. This procedure is repeated for each data object. The algorithm is composed of the following steps: [16]

1. Place K point into the space represented by the objects that are being clustered. These points represent initial group centroids;
2. Assign each object to the group that has the closest centroids;
3. When all objects have been assigned, recalculate the positions of the K centroids;
4. Repeat steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

Here, two parameters have to be decided in terms of mapping general K-means to our problems. One is the centroid. The centroid is the ideal condition for every water quality parameter to be in the safe range. The other is the closeness/proximity metrics. This is the standard deviation from the centroid, i.e., the ideal condition.

4.2.2. Calculation of Overall Water Quality Using Spider Index

The Spider Index is calculated by using a spider diagram. The quality of each parameter is calculated aggregated by year and station using the formula:

$$\text{Quality of Param} = \frac{\text{\#Observation Params in idea Range}}{\text{\#Total Observations}} \times 100 \tag{8}$$

Quality can be determined yearly, quarterly, monthly, or daily using this index. To calculate the overall quality index of the station, we can use following method:

- Mean of Parameter Quality: calculate the average of all the quality parameters to find the overall quality parameter, assuming all the parameters affect water quality equally. The larger the average, the better the quality.
- Area of Polygon: the area of polygon could be calculated to find the overall quality. A bigger area means better quality.

5. The Application of Platform

Here, we analyze the water quality of the San Francisco Bay area by using data from the sources: COAST and USGS. In our system, the water quality data of various regions of San Francisco Bay area are extracted by selecting the time period of the data [23]. Then we can analyze the data using different big data models and acquire some useful insights from the information. The system can better aid us to find the cause and curb down the pollutants in water.

5.1. Data Collection

Data Sources

USGS and CeNCOOS are the two data sources on water in San Francisco Bay. The station locations of USGS and CeNCOOs are shown in Figure 6. To aggregate the information of both the sources, we establish a mapping from CeNCOOS stations to USGS stations, which is shown in Table 5.

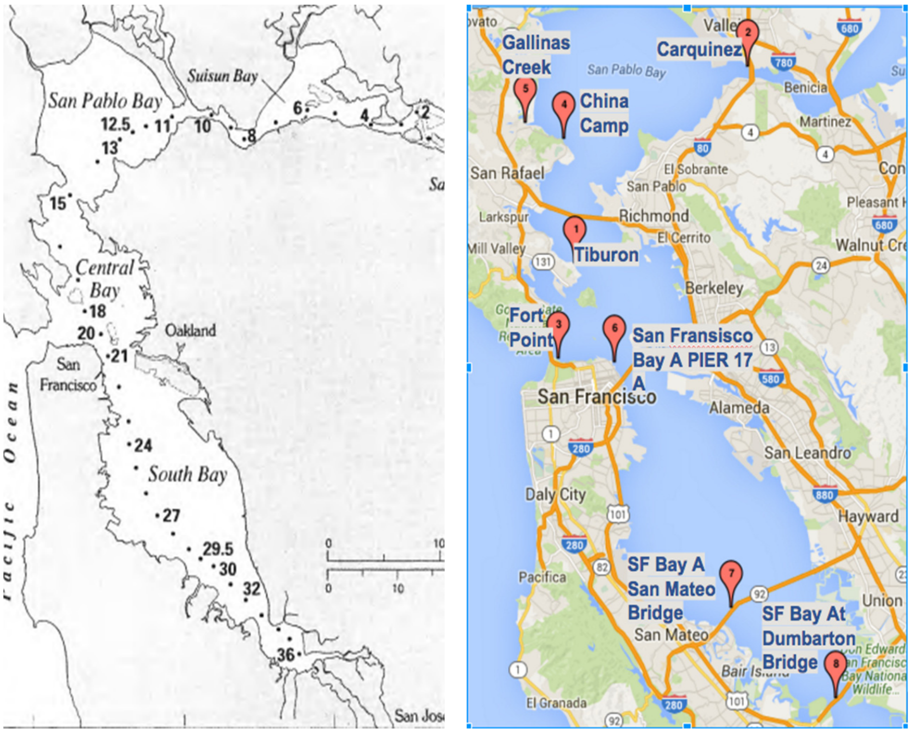


Figure 6. Station locations.

Table 5. Station Mapping.

CeNCOOS	Gallinas Creek	Dumbarton	San Mateo	Pier 17/Fort Point	Alcatraz	China Camp
USGS	13, 12.5, 11	36, 29.5, 30, 32	27, 24	21, 20	18	15

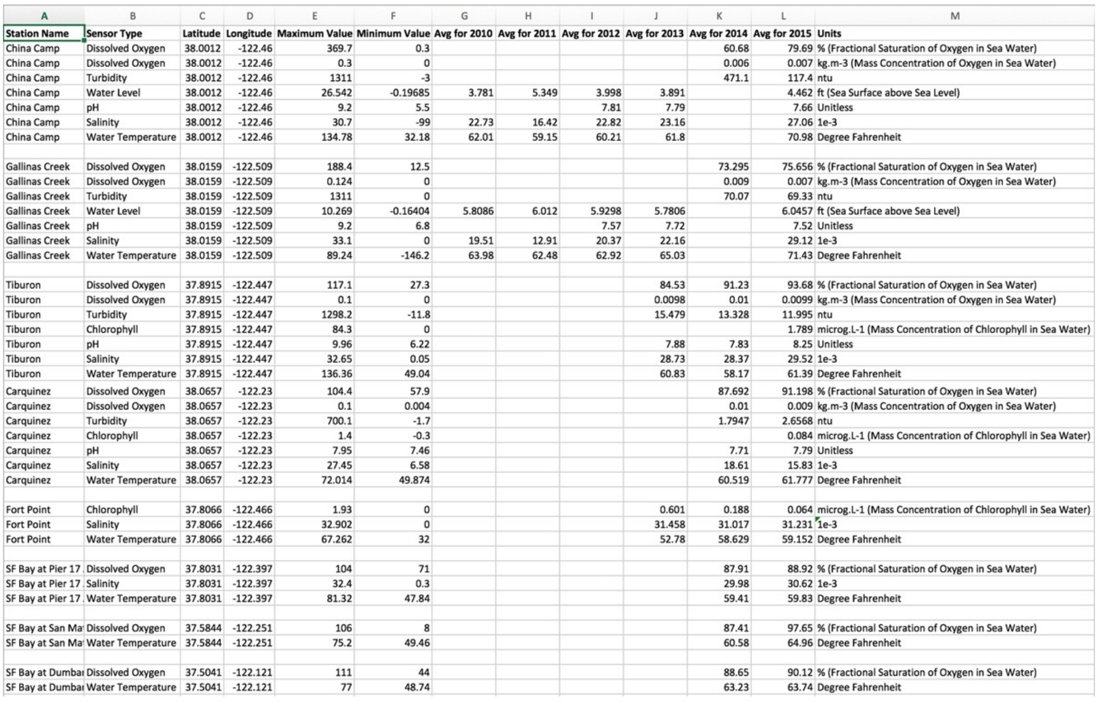
The stations in CeNCOOS and USGS collect the following water quality parameters: chlorophyll, dissolved oxygen, pH, salinity, turbidity, water temperature, and water level.

5.2. Data Extraction and Validation

Data Extraction

Figure 7 shows the sensor information file. This file contains the Station Name, Sensor Type, Latitude, Longitude, Maximum Value, Minimum Value, Average Values and Units.

The data extraction is automated using a BASH script, which will run periodically and then download and clean new data before pushing it into the database for use by the Web dashboard and charts. Data cleaning is also governed by the BASH script after data collection and is done in multiple iterations. The first iteration removes garbage values. These values could be any value that does not satisfy the variable type. The second iteration removes the out-of-range values. The third iteration removes all the duplicate rows. Finally, the redundant data are removed, such as sensor name and station name.

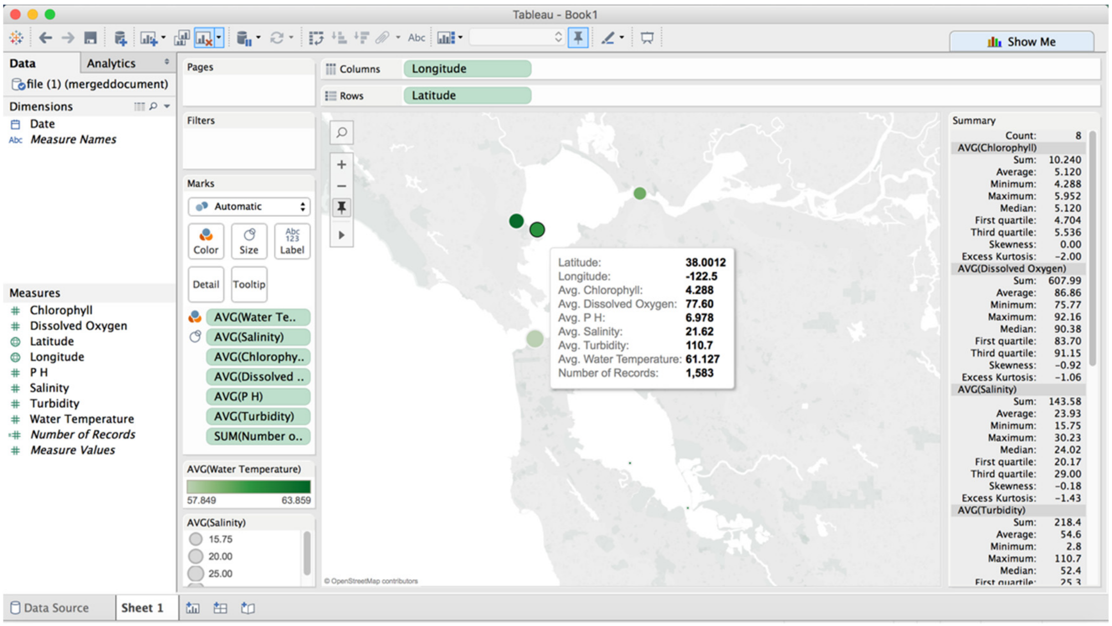


A	B	C	D	E	F	G	H	I	J	K	L	M
Station Name	**Sensor Type**	**Latitude**	**Longitude**	**Maximum Value**	**Minimum Value**	**Avg for 2010**	**Avg for 2011**	**Avg for 2012**	**Avg for 2013**	**Avg for 2014**	**Avg for 2015**	**Units**
China Camp	Dissolved Oxygen	38.0012	-122.46	369.7	0.3					60.68	79.69	% (Fractional Saturation of Oxygen in Sea Water)
China Camp	Dissolved Oxygen	38.0012	-122.46	0.3	0					0.006	0.007	kg.m-3 (Mass Concentration of Oxygen in Sea Water)
China Camp	Turbidity	38.0012	-122.46	1311	-3					471.1	117.4	ntu
China Camp	Water Level	38.0012	-122.46	26.542	-0.19685	3.781	5.349	3.998	3.891		4.462	ft (Sea Surface above Sea Level)
China Camp	pH	38.0012	-122.46	9.2	5.5			7.81	7.79		7.66	Unitless
China Camp	Salinity	38.0012	-122.46	30.7	-99	22.73	16.42	22.82	23.16		27.06	1e-3
China Camp	Water Temperature	38.0012	-122.46	134.78	32.18	62.01	59.15	60.21	61.8		70.98	Degree Fahrenheit
Gallinas Creek	Dissolved Oxygen	38.0159	-122.509	188.4	12.5					73.295	75.656	% (Fractional Saturation of Oxygen in Sea Water)
Gallinas Creek	Dissolved Oxygen	38.0159	-122.509	0.124	0					0.009	0.007	kg.m-3 (Mass Concentration of Oxygen in Sea Water)
Gallinas Creek	Turbidity	38.0159	-122.509	1311	0					70.07	69.33	ntu
Gallinas Creek	Water Level	38.0159	-122.509	10.269	-0.16404	5.8086	6.012	5.9298	5.7806		6.0457	ft (Sea Surface above Sea Level)
Gallinas Creek	pH	38.0159	-122.509	9.2	6.8			7.57	7.72		7.52	Unitless
Gallinas Creek	Salinity	38.0159	-122.509	33.1	0	19.51	12.91	20.37	22.16		29.12	1e-3
Gallinas Creek	Water Temperature	38.0159	-122.509	89.24	-146.2	63.98	62.48	62.92	65.03		71.43	Degree Fahrenheit
Tiburon	Dissolved Oxygen	37.8915	-122.447	117.1	27.3				84.53	91.23	93.68	% (Fractional Saturation of Oxygen in Sea Water)
Tiburon	Dissolved Oxygen	37.8915	-122.447	0.1	0				0.0098	0.01	0.0099	kg.m-3 (Mass Concentration of Oxygen in Sea Water)
Tiburon	Turbidity	37.8915	-122.447	1298.2	-11.8				15.479	13.328	11.995	ntu
Tiburon	Chlorophyll	37.8915	-122.447	84.3	0						1.789	microg.L-1 (Mass Concentration of Chlorophyll in Sea Water)
Tiburon	pH	37.8915	-122.447	9.96	6.22				7.88	7.83	8.25	Unitless
Tiburon	Salinity	37.8915	-122.447	32.65	0.05				28.73	28.37	29.52	1e-3
Tiburon	Water Temperature	37.8915	-122.447	136.36	49.04				60.83	58.17	61.39	Degree Fahrenheit
Carquinez	Dissolved Oxygen	38.0657	-122.23	104.4	57.9					87.692	91.198	% (Fractional Saturation of Oxygen in Sea Water)
Carquinez	Dissolved Oxygen	38.0657	-122.23	0.1	0.004					0.01	0.009	kg.m-3 (Mass Concentration of Oxygen in Sea Water)
Carquinez	Turbidity	38.0657	-122.23	700.1	-1.7					1.7947	2.6568	ntu
Carquinez	Chlorophyll	38.0657	-122.23	1.4	-0.3						0.084	microg.L-1 (Mass Concentration of Chlorophyll in Sea Water)
Carquinez	pH	38.0657	-122.23	7.95	7.46					7.71	7.79	Unitless
Carquinez	Salinity	38.0657	-122.23	27.45	6.58					18.61	15.83	1e-3
Carquinez	Water Temperature	38.0657	-122.23	72.014	49.874					60.519	61.777	Degree Fahrenheit
Fort Point	Chlorophyll	37.8066	-122.466	1.93	0				0.601	0.188	0.064	microg.L-1 (Mass Concentration of Chlorophyll in Sea Water)
Fort Point	Salinity	37.8066	-122.466	32.902	0				31.458	31.017	31.231	1e-3
Fort Point	Water Temperature	37.8066	-122.466	67.262	32				52.78	58.629	59.152	Degree Fahrenheit
SF Bay at Pier 17	Dissolved Oxygen	37.8031	-122.397	104	71					87.91	88.92	% (Fractional Saturation of Oxygen in Sea Water)
SF Bay at Pier 17	Salinity	37.8031	-122.397	32.4	0.3					29.98	30.62	1e-3
SF Bay at Pier 17	Water Temperature	37.8031	-122.397	81.32	47.84					59.41	59.83	Degree Fahrenheit
SF Bay at San Ma	Dissolved Oxygen	37.5844	-122.251	106	8					87.41	97.65	% (Fractional Saturation of Oxygen in Sea Water)
SF Bay at San Ma	Water Temperature	37.5844	-122.251	75.2	49.46					60.58	64.96	Degree Fahrenheit
SF Bay at Dumbar	Dissolved Oxygen	37.5041	-122.121	111	44					88.65	90.12	% (Fractional Saturation of Oxygen in Sea Water)
SF Bay at Dumbar	Water Temperature	37.5041	-122.121	77	48.74					63.23	63.74	Degree Fahrenheit

Figure 7. Sensor data file information.

5.3. Data Validation

After cleaning the newly imported data, we use Tableau Desktop 9.3 to validate data collected from CeNCOOS. We used the bubble diagram on a map and the XY Coordinate in Tableau to display the information from stations. An example of the validation results in our system is shown in Figure 8. The bubbles on the map are showing their locations, average values of all their sensors and the number of records. The size and the color of the bubbles are determined by the parameter values of water temperature and salinity. The detailed information on each parameter, such as number of records and maximum/minimum value, is shown on the right-hand side of Figure 8.



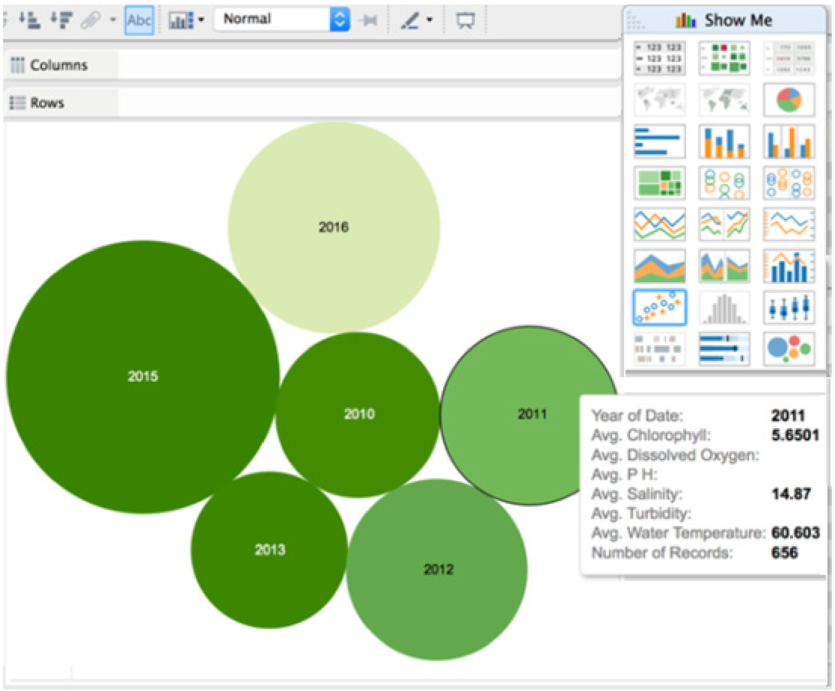
Latitude: 38.0012
Longitude: -122.5
Avg. Chlorophyll: 4.288
Avg. Dissolved Oxygen: 77.60
Avg. P H: 6.978
Avg. Salinity: 21.62
Avg. Turbidity: 110.7
Avg. Water Temperature: 61.127
Number of Records: 1,583

Summary
Count: 8
AVG(Chlorophyll)
Sum: 10.240
Average: 5.120
Minimum: 4.288
Maximum: 5.952
Median: 5.120
First quartile: 4.704
Third quartile: 5.536
Skewness: 0.00
Excess Kurtosis: −2.00
AVG(Dissolved Oxygen)
Sum: 607.99
Average: 86.86
Minimum: 75.77
Maximum: 92.16
Median: 90.38
First quartile: 83.70
Third quartile: 91.15
Skewness: −0.92
Excess Kurtosis: −1.06
AVG(Salinity)
Sum: 143.58
Average: 23.93
Minimum: 15.75
Maximum: 30.23
Median: 24.02
First quartile: 20.17
Third quartile: 29.00
Skewness: −0.18
Excess Kurtosis: −1.43
AVG(Turbidity)
Sum: 218.4
Average: 54.6
Minimum: 2.8
Maximum: 110.7
Median: 52.4
First quartile: 25.3

Figure 8. The validation results in our system.

5.4. Data Storage

After data validation, the data are stored in the system database. We also use a bubble view to show each year of data, which is shown in Figure 9. The size of the bubble denotes the counts of records available for that year and the color represents the average water temperature for that year. By double clicking the bubble, we can obtain the statistics information of all parameters in all stations ordered by year. The Gantt view is used to illustrate the entire dataset in our system, which is shown in Figure 10. It is a good way to recognize variation in data.



Year of Date: 2011
Avg. Chlorophyll: 5.6501
Avg. Dissolved Oxygen:
Avg. P H:
Avg. Salinity: 14.87
Avg. Turbidity:
Avg. Water Temperature: 60.603
Number of Records: 656

Figure 9. Bubble view for data of each year.

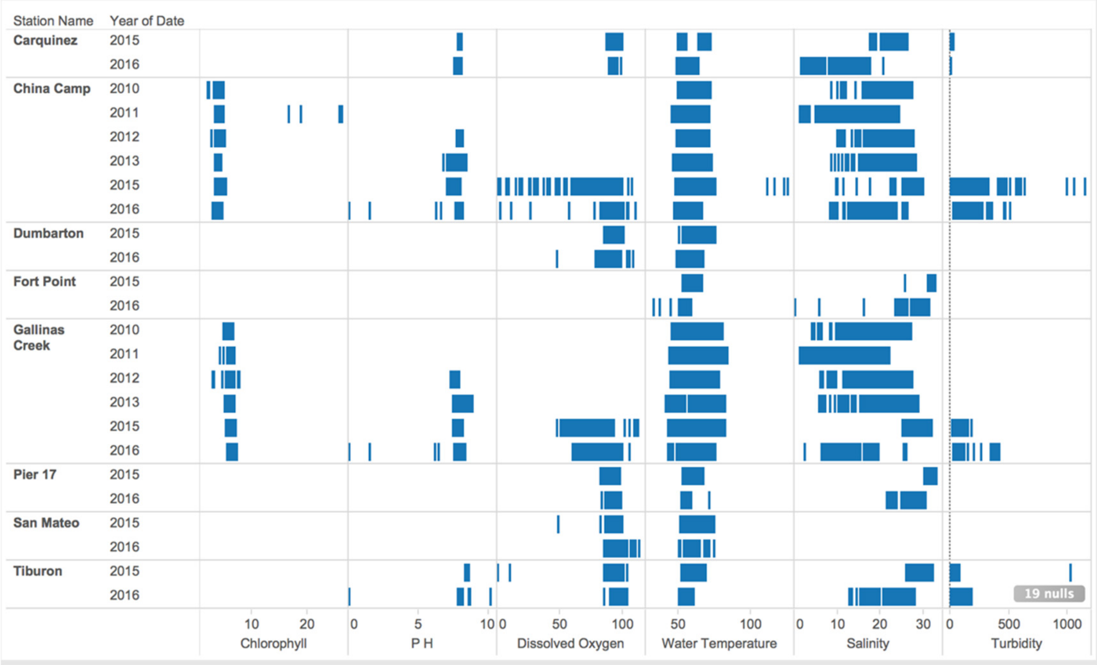


Figure 10. Gantt view of the entire data set.

5.5. Data Presentation

A friendly UI in our system is provided to check the sensor location, status, and its data, which is shown in Figure 11. Clicking on a sensor brings up the data available for that sensor, which is shown in Figure 12.

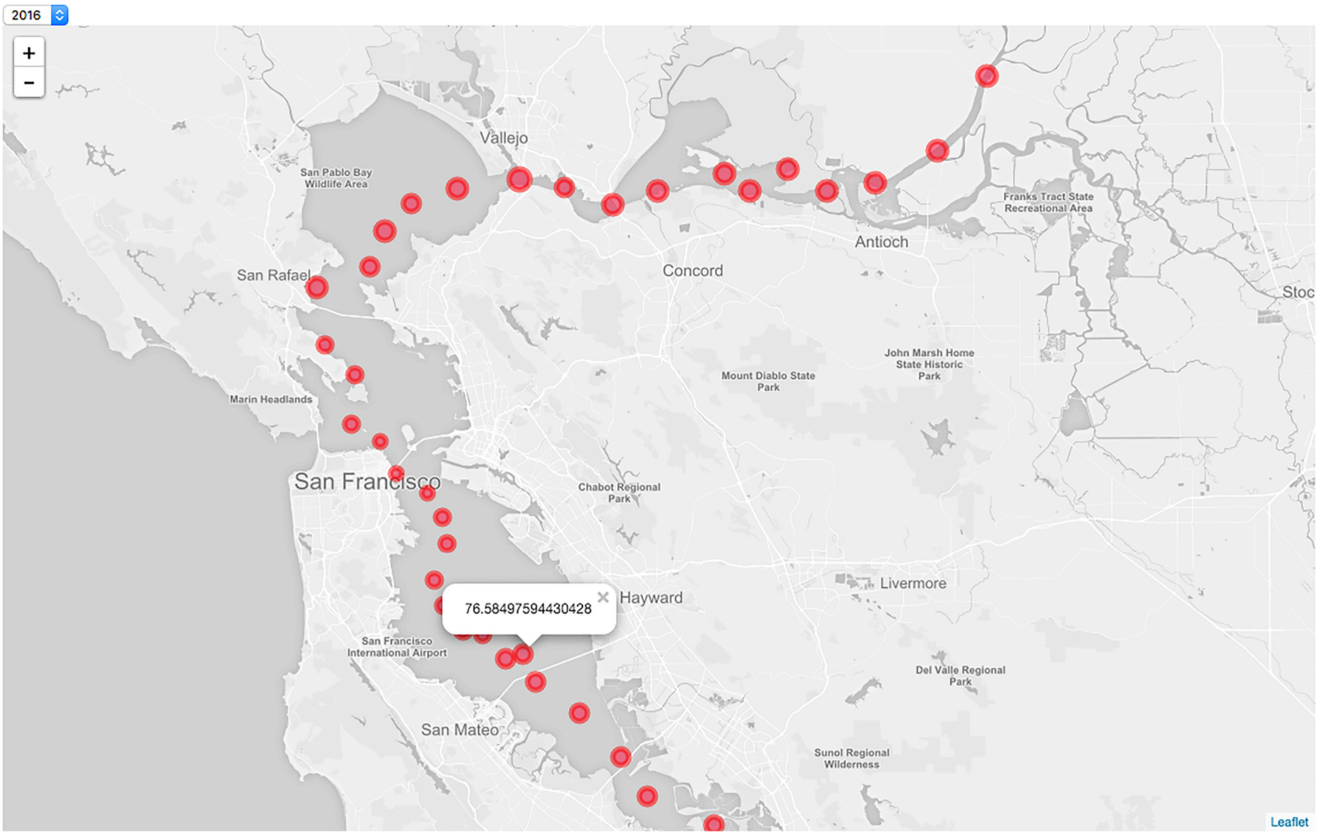


Figure 11. Location of sensors.

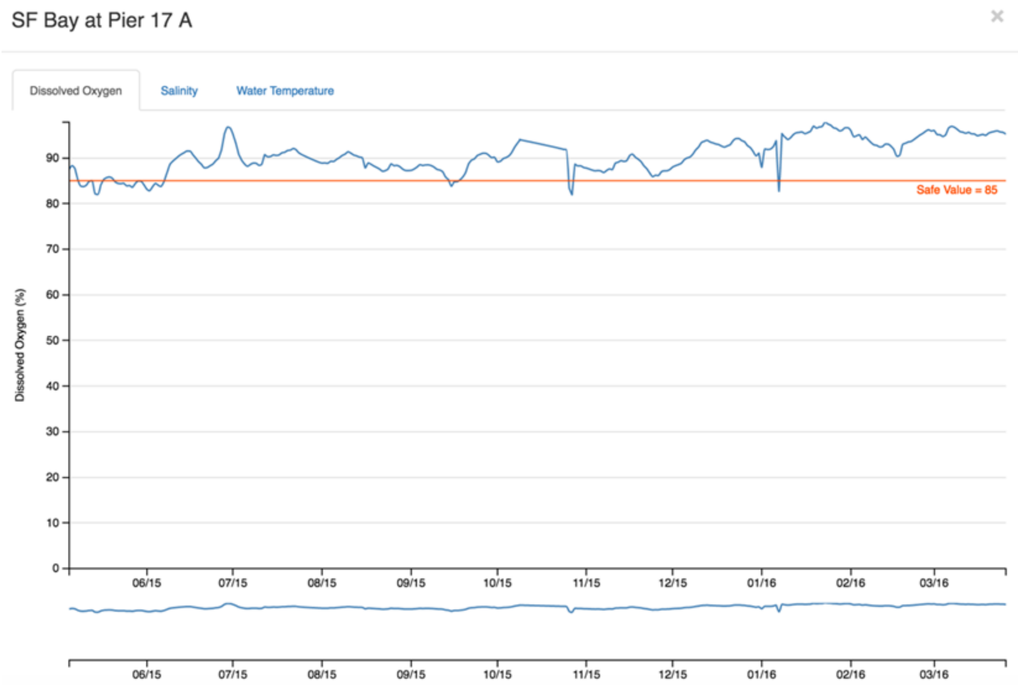

SF Bay at Pier 17 A

Figure 12. The data records in the sensor.

Based on the analysis model, the water quality can be evaluated by different approaches. Figure 13 shows the result according to the Water Quality Index. The result also can be presented by the Sankey chart, which is shown in Figure 14.

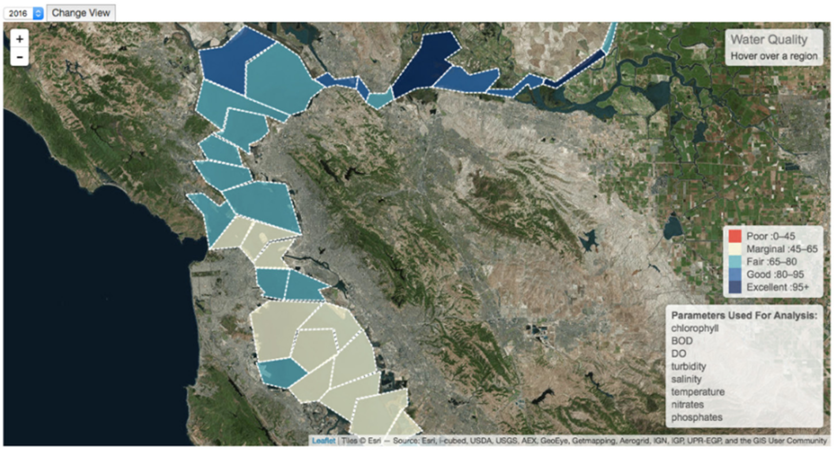


Figure 13. Water Quality Index in the San Francisco Bay.

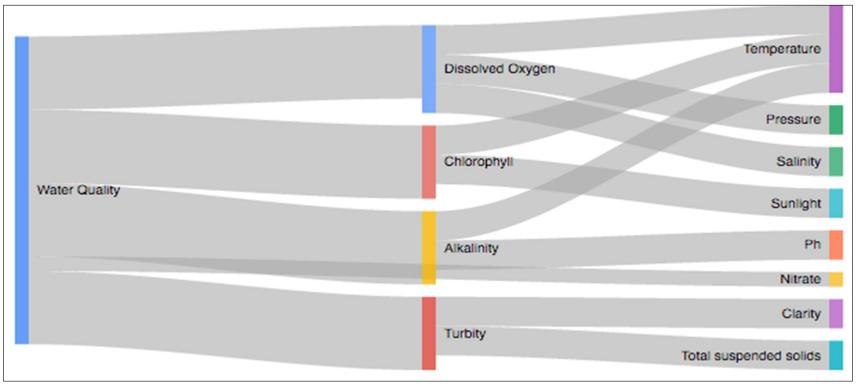


Figure 14. Sankey Chart displaying water quality dependencies.

Moreover, we can also obtain the analysis results by some methods used in big data analysis. Figures 15 and 16 show the results by using spider index. Figures 17 and 18 show the K-means classification results of the water quality by bubble view.

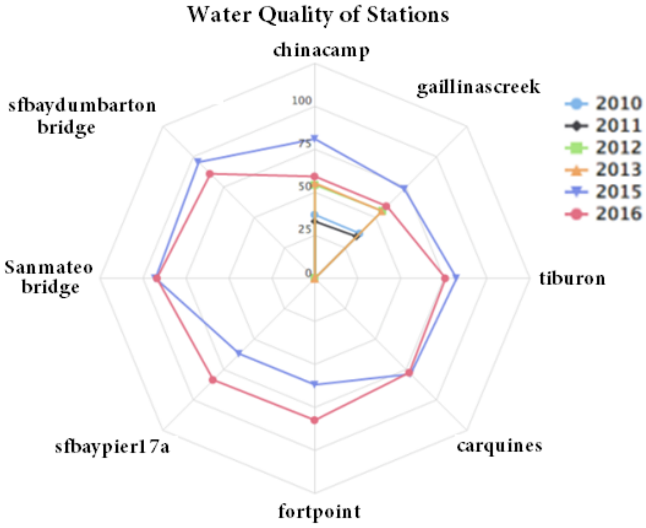


Figure 15. Spider diagram showing water quality of a specific station.

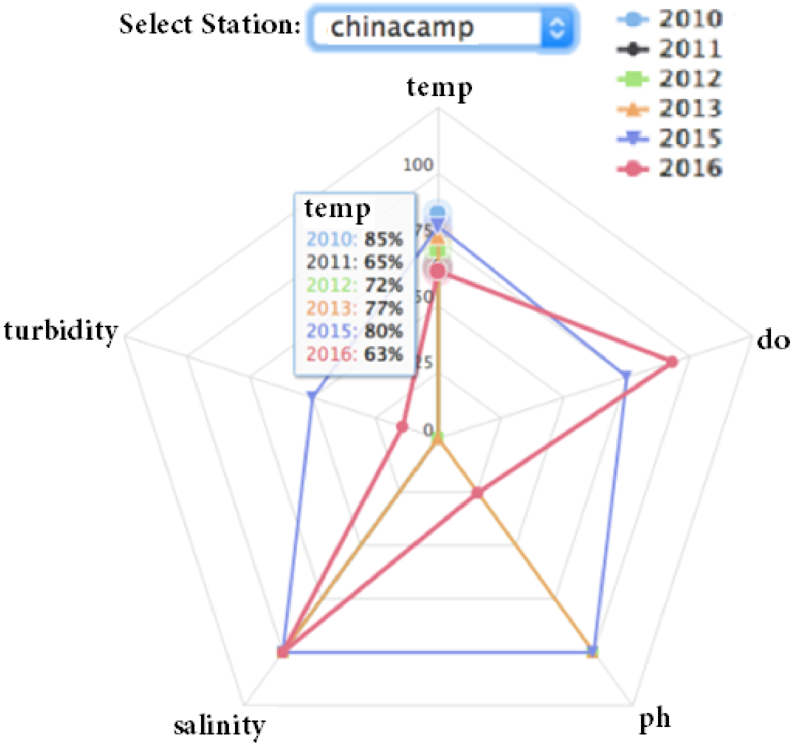


Figure 16. Spider diagram showing water quality of all stations.

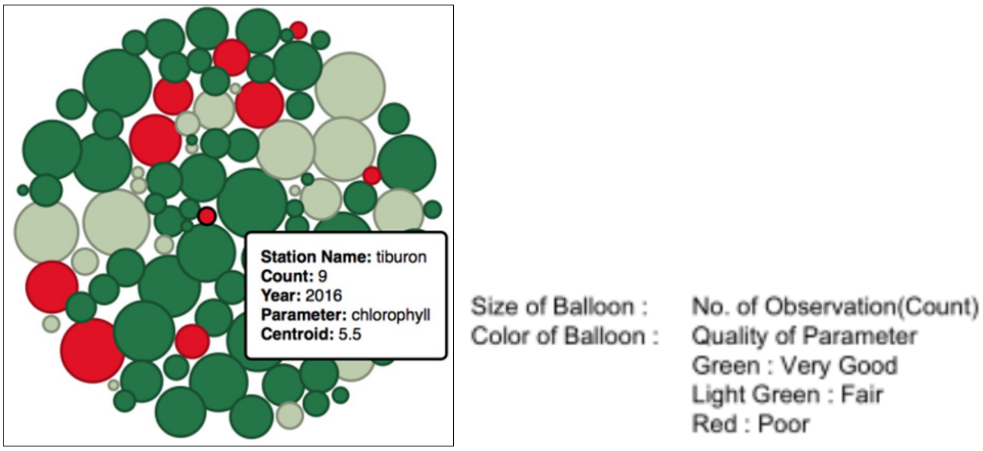


Figure 17. K-Means classification using a bubble diagram.

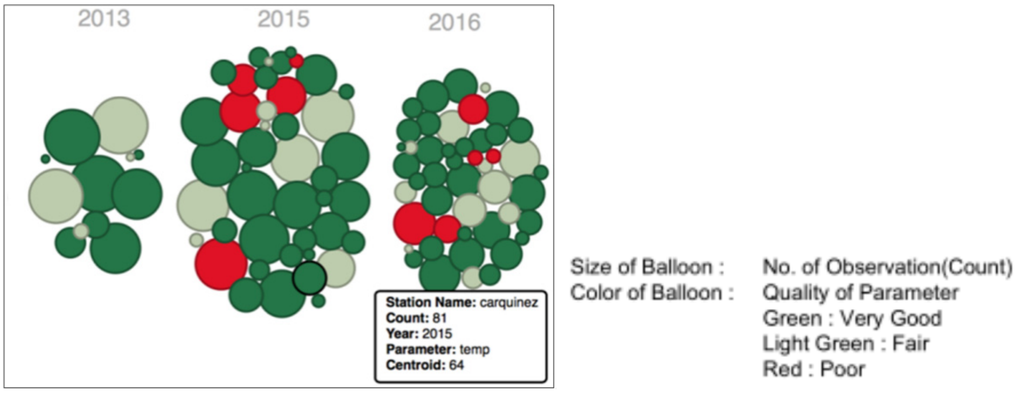


Figure 18. Bubble diagram classified for different years.

6. Conclusions

In this paper, we first summarized the water quality index for different uses. Then, we developed a Web-based platform, which provides an open interface to access the water quality data from the sensors. Moreover, both the traditional approaches and big data methods are integrated in this platform. With the aid of the platform, we can analyze the water quality based on not only the traditional methods but also based on big data analysis. Some popular methodologies and technologies used to develop our platform are Cassandra for our big data database, Google High Chart for coloring region mapping, and spider diagrams. The users can better understand and use the analysis results. Therefore, the platform can be a good tool that we can use to analyze water quality.

Author Contributions: Data curation, X.H.; Formal analysis, X.H.; Funding acquisition, H.S.; Methodology, H.S.; Project administration, H.H. and J.G.; Resources, J.G.; Writing—original draft, X.H. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by the project "CERN long-term observation data mining and annual data report" (Grant KFJ-SW-YW043) and "Community structure of macrobenthos and microbial diversity in sediments of Northwest Lake and Guizi Lake" (Grant 3015/08210056).

Institutional Review Board Statement: This study does not involve human or animal research.

Data Availability Statement: Data is collected from sensors installed across the San Francisco Bay by Organizations like CeNCOOS (Central and Northern California Ocean Observing System) and USGS's NWIS (U.S. Geological Survey's National Water Information System).

Acknowledgments: The authors would like to thank Dolly Gupta, Vincent Poon, Sonam Saini, Madhav Vasudev, Yong Wang and Huiming Cao for their contribution of in designing and developing the water evaluation platform.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Diersing, N.; Nancy, F. *Water Quality: Frequently Asked Questions*; Florida National Marine Sanctuary: Key West, FL, USA, 2009.
2. Abbasi, T.; Abbasi, S.A. *Water Quality Indices*; Elsevier: Amsterdam, The Netherlands, 2012.
3. He, T.; Lu, Y.; Cui, Y.; Luo, Y.; Wang, M.; Meng, W.; Zhang, K.; Zhao, F. Detecting gradual and abrupt changes in water quality time series in response to regional payment programs for watershed services in an agricultural area. *J. Hydrol.* **2015**, *525*, 457–471. [CrossRef]
4. Lobato, T.C.; Hauser-Davis, R.A.; Oliveira, T.F.; Silveira, A.M.; Silva, H.A.N.; Tavares, M.R.M.; Saraiva, A.C.F. Construction of a novel water quality index and quality indicator for reservoir water quality evaluation: A case study in the Amazon region. *J. Hydrol.* **2015**, *522*, 674–683. [CrossRef]
5. Newton, A.; Mudge, S.M. Lagoon-sea exchanges, nutrient dynamics and water quality management of the Ria Formosa (Por-tugal). *Estuar. Coast. Shelf Sci.* **2005**, *62*, 405–414. [CrossRef]
6. Newton, A.; Mudge, S.M. Temperature and salinity regimes in a shallow, mesotidal lagoon, the Ria Formosa, Portugal. *Estuar. Coast. Shelf Sci.* **2003**, *57*, 73–85. [CrossRef]
7. Lattemann, S.; Höpner, T. Environmental impact and impact assessment of seawater desalination. *Desalination* **2008**, *220*, 1–15. [CrossRef]
8. Alizadeh, M.J.; Kavianpour M, R. Development of wavelet-ANN models to predict water quality parameters in Hilo Bay, Pacific Ocean. *Mar. Pollut. Bull.* **2015**, *98*, 171–178. [CrossRef] [PubMed]
9. Lima, E.; Legey, L.F.L. Water Quality Restoration in Rio de Janeiro: From a Piecemeal to a Systems Approach. *J. Environ. Dev.* **2010**, *19*, 375–396. [CrossRef]
10. Shrestha, S.; Kazama, F. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environ. Model. Softw.* **2007**, *22*, 464–475. [CrossRef]
11. Sridhar, R.; Thangaradjou, T.; Kumar, S.S.; Kannan, L. Water quality and phytoplankton characteristics in the Palk Bay, southeast coast of India. *J. Environ. Biol.* **2006**, *27*, 561–566. [PubMed]
12. Vega, M.; Pardo, R.; Barrado, E.; Debán, L. Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Res.* **1998**, *32*, 3581–3592. [CrossRef]
13. Coskun, H.G.; Tanik, A.; Alganci, U.; Cigizoglu, H.K. Determination of environmental quality of a drinking water reservoir by remote sensing, GIS and regression analysis. *Water Air Soil Pollut.* **2008**, *194*, 275–285. [CrossRef]
14. Beck, M.W.; Hagy, J.D., III. Adaptation of a Weighted Regression Approach to Evaluate Water Quality Trends in an Estuary. *Environ. Modeling Assess.* **2015**, *20*, 637–655. [CrossRef]

15. Chenini, I.; Khemiri, S. Evaluation of ground water quality using multiple linear regression and structural equation modeling. *Int. J. Environ. Sci. Technol.* **2009**, *6*, 509–519. [CrossRef]
16. Modaresi, F.; Araghinejad, S. A Comparative Assessment of Support Vector Machines, Probabilistic Neural Networks, and K-Nearest Neighbor Algorithms for Water Quality Classification. *Water Resour. Manag.* **2014**, *28*, 4095–4111. [CrossRef]
17. Areerachakul, S.; Sanguansintukul, S. *Clustering Analysis of Water Quality for Canals in Bangkok, Thailand*; International Conference on Computational Science and Its Applications; Springer: Berlin/Heidelberg, Germany, 2010; pp. 215–227.
18. Aryafar, A.; Gholami, R.; Rooki, R.; Ardejani, F.D. Heavy metal pollution assessment using support vector machine in the Shur River, Sarcheshmeh copper mine, Iran. *Environ. Earth Sci.* **2012**, *67*, 1191–1199. [CrossRef]
19. Liu, M.; Lu, J. Support vector machine—an alternative to artificial neuron network for water quality forecasting in an agricultural nonpoint source polluted river? *Environ. Sci. Pollut. Res.* **2014**, *21*, 11036–11053. [CrossRef] [PubMed]
20. Noori, R.; Karbassi, A.; Ashrafi, K.; Ardestani, M.; Mehrdadi, N.; Bidhendi, G.-R.N. Active and online prediction of BOD5 in river systems using reduced-order support vector machine. *Environ. Earth Sci.* **2011**, *67*, 141–149. [CrossRef]
21. Kaur, M.; Das, S.K.; Sarma, K. Water quality assessment of Tal Chhapar Wildlife Sanctuary using water quality index (CCME *WQI*). *Acta Ecol. Sin.* **2021**. [CrossRef]
22. Masoud, A.M.; Ali, M.H. Coupled multivariate statistical analysis and *WQI* approaches for groundwater quality assessment in Wadi El-Assiuty downstream area, Eastern Desert, Egypt. *J. Afr. Earth Sci.* **2020**, *172*, 103982. [CrossRef]
23. Dolly, G.; Vincent, P.; Sonam, S.; Madhav, V. Big Data Based Water Quality Evaluation for San Francisco Bay Based on COAST Data. A Project Report Presented to The Faculty of the Computer Engineering Department. Master's Thesis, San Jose State University, San Jose, CA, USA.