Water Quality Database Management in PA

12741 Project Report

Zhen Ji
Civil and Environmental Engineering
Carnegie Mellon University
Pittsburgh, PA
zhenji@andrew.cmu.edu

Yuhang Liang
Civil and Environmental Engineering
Carnegie Mellon University
Pittsburgh, PA
yuhangl@andrew.cmu.edu

Lin Lyu
Civil and Environmental Engineering
Carnegie Mellon University
Pittsburgh, PA
linlyu@andrew.cmu.edu

1 Introduction

1.1 Motivation

Water pollution occurs in various parts of the world and substandard water endangers the health of people seriously. Therefore, water quality testing and management of water quality data are extremely important. In this project, we will design a database about water quality in Pennsylvania Allegheny County and do some data analysis to help relevant departments to make decisions.

1.2 Data Sources

The data used is from Water Quality Portal (WQP), which integrates publicly available water quality data from the USGS National Water Information System (NWIS). The data sources are from: https://www.waterqualitydata.us/portal/. The database is redundant. It has many blank columns, partial dependencies and transitive dependencies.

1.3 Objective

The first objective is to build a database about water quality in Pennsylvania Allegheny County by analyzing the relations between entities, designing the E-R diagram to represent the entities, attributes and their relationships, doing database normalization and achieving it on MySQL.

Secondly, we would add two applications for the database so that we can quickly obtain the required data through the query command on mysql, such as finding when and where the concentration of phosphorus exceeded the water quality standard value and selecting the river location and the corresponding sampling time with substandard pH.

Thirdly, in data analysis part, we would conduct linear regression fitting on three types of data: dissolved oxygen, temperature, and pH and obtain a formula based on time.

2 Database Design

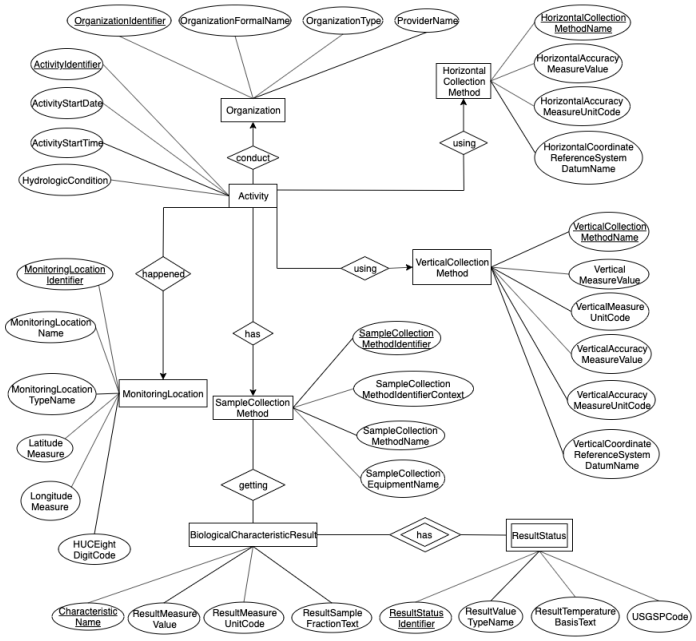

Figure 1: E-R diagram of water quality database

From Figure1, we can see the database has seven entities: organization, activity, monitoring location, sample collection method, biological characteristic result, horizontal collection method, vertical collection method and result status. And each entity has several attributes.

As for the relationship between entities, one organization can conduct several activities, so the relationship is “one to many”. Several activities can happen in one monitoring location and the relationship is “many to one”. The relationship between activity and sample collection method is “many to one”. The relationship between activity and horizontal collection method is “many to one”. The relationship between activity and vertical collection method is “many to one”. The relationship between activity and sample collection method is “many to one”. The relationship between sample collection method and biological characteristic result is “many to many”. The last relationship in this E-R diagram is “many to many” which is between biological characteristic result and result status. And it is noted that result status is a weak entity

because its existence has to rely on the biological characteristic result.

3 Database Normalization

The original database we downloaded includes activity.xlsx, organization.xlsx, biologicalresult.xlsx and station.xlsx. Organization.xlsx records information about the organization that collects and detects water samples, including organization identifier, normal name of the organization, and other related Information. Activity.xlsx records the sampling activities conducted by the organization, including organization identifier, activity identifier, and sampling site information. Station.xlsx records the sampling site number, the latitude and longitude of the sampling site, and other information. Biologicalresult.xlsx records sampling organization, sampling method, sampling indicators and its values.

Therefore, there are many blank columns, partial dependencies and transitive dependencies in the databases we obtained. In order to minimize database redundancy, it is necessary to do database normalization.

3.1 First Normalization Form

In the databases, many blank columns and duplicate columns make the database have many invalid parts. We derived and removed these columns in the database by using Mysql. After that, each column of the processed database contains atomic values, which meets the requirements of 1NF.

3.2 Second Normalization Form

Based on 1NF, we started to extract the partial dependencies existing in the database. Take biologicalresult.csv as an example, there are organizationIdentifer and organizationNormalName in the database. Actually, the identifier determined the normal name. The existence of this dependency makes the database expansion meaningless to some extent, because the data related to the organization can be retrieved in the organization.csv file through the identifier. By deleting and creating new databases of new entities, we eliminated the partial dependencies in the database and formed 2NF.

3.3 Boyce-Codd Normal Form and 3NF

Based on 2NF, we deleted the transitive dependencies in the database. Take biologicalresult.csv as an example, there is a sampleCollectionMethod entity and its attributes in the database, and there is a function dependency between the method entity and the identifier for water sample. This creates a transitive decency. We eliminated this kind of decencies during this normalization and formed 3NF. Since the primary key of our database entity consists of only one attribute, and for any dependency A→B, A is a super key. Therefore, we obtained BCNF through this normalization except biologicalresult.csv, which is a 3NF. Because in this database, we have to use three keys to identify a specific sample result, ActivityIdentifier, MonitoringLocationIdentifier and CharacteristicName.

4 Database Application

Based on the normalized databases, we wrote queries to extract data on Mysql. Our goal is to find substandard river water samples in the database and their sampling time. We mainly accomplished two data application goals.

4.1 Finding Potential Eutrophication

The first goal is to find when and where the concentration of phosphorus exceeded, which has a great relationship with eutrophication of water. By looking for water quality standards, if the total phosphorus concentration exceeds 0.1 mg / L, the river water quality cannot meet the class IV standard. By querying on Mysql, we obtained the activitIdentifier and locationIdentifer of the over-standard water sample through biologicalresult.csv, and connected to activity.csv and station.csv through two identifiers to obtain the corresponding sampling time and sampling location. This database can intuitively reflect over-standard rivers and over-standard events.

MonitoringLocationName	ActivityStartD...	CharacteristicName	ResultMeasureValue	ResultMeasure/Measur...
Monongahela River at Elizabeth, PA	2010-12-07	Phosphorus	0.125	mg/l as P
Monongahela River at Elizabeth, PA	2017-02-07	Phosphorus	0.25	mg/l as P
Monongahela River at Elizabeth, PA	2017-03-16	Phosphorus	0.99	mg/l as P
Monongahela River at Elizabeth, PA	2017-11-07	Phosphorus	0.16	mg/l as P
Monongahela River at Elizabeth, PA	2018-02-12	Phosphorus	0.13	mg/l as P
Youghiogheny River at Sutersville, PA	2009-01-07	Phosphorus	0.164	mg/l as P
Youghiogheny River at Sutersville, PA	2012-05-10	Phosphorus	0.113	mg/l as P
Youghiogheny River at Sutersville, PA	2013-06-26	Phosphorus	0.627	mg/l as P
Youghiogheny River at Sutersville, PA	2018-02-12	Phosphorus	0.16	mg/l as P
Youghiogheny River at Sutersville, PA	2018-06-11	Phosphorus	0.31	mg/l as P
Youghiogheny River at Sutersville, PA	2019-02-25	Phosphorus	0.12	mg/l as P
Youghiogheny River at Sutersville, PA	2019-06-04	Phosphorus	0.34	mg/l as P
Ohio River at Sewickley, PA	2010-04-08	Phosphorus	0.575	mg/l as P
Ohio River at Sewickley, PA	2014-08-27	Phosphorus	0.11	mg/l as P
Ohio River at Sewickley, PA	2019-02-21	Phosphorus	0.13	mg/l as P
Ohio River at Sewickley, PA	2019-05-08	Phosphorus	0.28	mg/l as P

Figure 2: Database Substandard Database for TP

4.2 Finding Substandard Water Quality

pH is an important indicator of water quality because it affects many other water quality indicators and it is of great significance for the growth of aquatic organisms. We used LEFT JOIN (this is because some monitoring locations do not have official names) and INNER JOIN to connect the three databases of station.csv, biologicalresult.csv, and activityall.csv, and finally selected the river location and the corresponding sampling time with substandard pH.

MonitoringLocationName	ActivityStartD...	CharacteristicName	ResultMeasureValue	ResultMeasure/Measur...
Allegheny River at 9th St Bridge at Pittsburgh, PA	2009-06-18	pH	8.6	std units
Youghiogheny River at Sutersville, PA	2010-10-12	pH	8.6	std units
Youghiogheny River at Sutersville, PA	2010-10-12	pH	8.6	std units
Youghiogheny River at Sutersville, PA	2010-10-12	pH	8.6	std units
Youghiogheny River at Sutersville, PA	2010-10-12	pH	8.6	std units
Youghiogheny River at Sutersville, PA	2010-10-12	pH	8.6	std units
Youghiogheny River at Sutersville, PA	2010-10-12	pH	8.6	std units
Youghiogheny River at Sutersville, PA	2010-10-12	pH	8.6	std units
NULL	2009-07-07	pH	8.8	std units
NULL	2009-07-07	pH	8.9	std units
NULL	2009-07-07	pH	8.9	std units
NULL	2009-08-26	pH	8.6	std units
Turtle Creek at Wilmerding, PA	2009-08-26	pH	9.1	std units
Thompson Run at Turtle Creek, PA	2009-08-26	pH	8.7	std units

Figure 3: Substandard Database for pH

5 Data Analysis

For this step, we applied what we learned in time series. Basically, we want to get the relevant feature of several selected attributes

mainly from the original biological result table which records the value of many characteristics like dissolved oxygen, carbon dioxide, water temperature, pH value, etc. To complete the analysis, we picked dissolved oxygen, water temperature, pH as our analytical value.

5.1 Data Reorganization

The database we created so far can satisfy the application we mentioned before. However, the table which reflects the characteristic value isn't a normal time series form. Instead of having a relation where each time maps several characteristic values, original relation has activities (which is identified by the activity time) mapping one characteristic at a time, and only by several records can we get the values of the characteristics at a specific time. Also, different locations' data has been merged together which makes the analysis more complex.

ActivityIdentifier	Activity	Activity	Activity	Activity	ActivityS	Activity	CharacteristicName	ResultS
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Oxygen	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Enrofloxacin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Chloramphenicol	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Isochlortetracycline	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Isoepichlorotetracycline	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Epi-tetracycline	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Anhydroerthromycin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Ormetoprim	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Lomefloxacin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Ofloxacin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Ciprofloxacin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Roxithromycin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Lincomycin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Carbamazepine	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Azithromycin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Tetracycline	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Sulfathiazole	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Sulfachloropyridazine	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Sarafloxacin	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Doxycycline	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Trimethoprim	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Benzeneacetic acid, .alp	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Sulfamethazine	Dissolved
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Hydrogen ion	Total
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Gage height	
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	Stream flow, instantaneous	
nwispa.01.00900571	Sample-R	Water	Surface W	2009-03-2	12:40:00	EDT	pH	Total

Figure 4: Original Relation

So, we selected the values of dissolved oxygen, temperature and pH within the same location to form 3 tables which have each characteristic's value mapping time. And we also deleted some duplicates. After that, we combine the 3 values together by time order to form one single relation.

5.2 Linear Regression Process

In the analysis step, we used linear regression method to analyze the there attributes. We assumed that the dissolved oxygen is based on time and other two factors: pH and temperature.

$do_i = q_i + \gamma pH_i + \delta T_i + n_i$

$q_i = \alpha t_i + \beta$

(do: dissolved oxygen, q = actual do, T: water temperature)

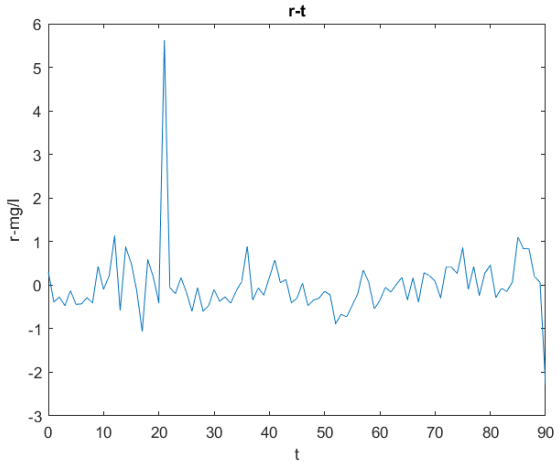

Figure 5: Residual plotting

After we got a model of the dissolved oxygen, we also used Chauvenet's criterion to filter outliers and moved them, after that we use the new data to yield a model, after 3 iterations we got a more accurate model.

The results are:

α=-0.0032(mg/(l*month));

β=12.0235(mg/l);

γ=0.2859(mg/l);

δ=-0.2112(mg/(l*degree C))

6 Discussion and Conclusion

In this project, we have finished a relatively complete data management process, including the collection of databases, the design and normalization of new databases, the application of databases, and data analysis. We downloaded some databases of surface water sampling results. Through the process of drawing E-R diagrams, we identified entities, their attributes and relationships. In the process of database normalization, we separated necessary databases to minimize database redundancy from the original database. And in the database application phase, we can quickly obtain the required data through the query command on mysql. Finally, in the data analysis part, we conducted linear regression fitting on three types of data: dissolved oxygen, temperature, and pH, and then used Chauvenet's criterion to filter outliers. The schema of our project includes normalized biologicalresult table, station table, activityall table, HorizontalCollectionMethod table, VerticalCollectionMethod table, etc.

REFERENCES

[1] EPA. CHAPTER 62-302 SURFACE WATER QUALITY STANDARDS. https://www.epa.gov. April 30, 2018