Mining spatial and temporal data to classify water quality: a case study

A. J. Seixas[1], S. L. P. L. Beatriz[2] & F. F. E. Nelson[2]
[1]*Eletronic Computing Center, Federal University of Rio de Janeiro, Brazil*
[2]*Civil Engineering Program, COPPE and Polythecnic School, Federal University of Rio de Janeiro, Brazil*

Abstract

Water is part of the heritage of the planet. Every continent, nation, region, city, and citizens are fully responsible for it. The environment and the future of our planet depend on the preservation of water and its cycles. Those should remain intact and work normally in order to ensure continuity of life on Earth. This balance depends, in particular, on the preservation of the seas and oceans where the cycles begin.

The water management requires a balance between the necessity of environmental protection and the need for economic order, health and social affairs. In Brazil, around 80% of the sewage is discharged into water bodies without treatment, in which 85% are domestic sewage and 15% industry sewerage. Therefore, the conditions of sanitation are poor in most Brazilian cities. This scenario is not very different in the City of Rio de Janeiro. In the waters of the Rodrigo de Freitas Lagoon, the problem of sewage release has mostly been solved, but there is still much to do. A better control over the quality of its waters is necessary, in contrast to the lack of public resources needed for this.

This work proposes to investigate the correlations of the spatial and temporal data that compose the set of pollutants of the Rodrigo de Freitas Lagoon. The main goal is to achieve a methodology for the classification of water quality, which may be used in other water bodies. For that, in the investigative process, several techniques of data mining will be used to group and classify the available data. The work includes several steps of knowledge discovery that will be implemented to reach the goals. Tasks such as cleaning, integration and data selection, as well as post-processing and visualization algorithms must be executed. Some preliminary results, already available, will be shown in this article.

Keywords: data mining, environment, classification, clustering, water quality, pollution.

WIT Transactions on Information and Communication Technologies, Vol 40, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/DATA080091

1 Introduction

Water is vital to the existence of all living organisms, but this feature is increasingly being threatened by the growth of human population, which requires more high-quality water for domestic purposes and economic activities. The use of water for domestic, agricultural production, mining, manufacturing, energy generation, coupled with the irregular deforestation and occupation of the soil, can lead to deterioration in the quality and quantity of water that not only impact the aquatic ecosystem (i.e., a collection of organisms that live in and interact together within an aquatic environment), but also the availability of water safe for human consumption. These aquatic environments are complex matrices that require careful use to ensure a self-sustaining ecosystem that works well in the future. Moreover, the maintenance of these environments requires a better understanding of connections between the properties of the ecosystem and the way in which human activities can change the inter-relationship between physical, chemical and biological processes, conducting the functioning of the ecosystem. Water, in the natural environment, contains many substances dissolved and particulate matter non-dissolved. Dissolved salts and minerals are necessary components in the water for a good quality because they help to maintain health and vitality of organisms that rely on this service of the ecosystem [1].

Water from natural sources almost always contains living organisms which are components of the biochemical cycles in aquatic ecosystems. However, some of these, particularly bacteria, parasitic worms, fungi, and viruses, that are present in the water can be harmful to humans.

Typically, the water quality is determined by comparing the physical and chemical characteristics of a sample with quality standards. Patterns of drinking water are set to provide clean and safe water thus are protecting human health. They are usually based on scientifically acceptable levels of toxicity to both humans and aquatic organisms. The standards for the protection of aquatic life are more difficult to define, mainly because the aquatic ecosystems vary enormously in space and time, in its composition and because the limits of an ecosystem rarely coincide with Territorial limits. Consequently, there is a movement among scientists and regulatory agencies to identify the natural conditions of the chemical characteristics that are not toxic to humans and animals and to use these as standards for the protection of aquatic life [2]. Other standards, such as those designed to ensure adequate quality for recreation activities, agricultural or industrial, define the limits for physical, chemical and biological water, necessary to undertake various activities safely.

2 Problem characterization

2.1 The Rodrigo de Freitas Lagoon

The Rodrigo de Freitas Lagoon, located in the South Zone of the City of Rio de Janeiro, has an extension of 2.2 km², depth average of 2.8 m and 7.8 km of

perimeter, with a volume of approximately 6.200.000m³. It binds to the sea by a channel, which has 800m in length and between 10 and 18 meters wide.

Figure 1: Rodrigo de Freitas Lagoon and its monitoring points.

Located between the south side of the Carioca Mountain and the sea, in an urban area of high population density in the city of Rio de Janeiro, the Rodrigo de Freitas Lagoon received disposal household for a long period and still receives accidentally, and is therefore in the process of eutrophication. Improvements undertaken by the government of Rio de Janeiro state in 2001, to expand the network collector, eliminated these disposals in clear days.

2.2 The FEEMA classification

The State Engineering Foundation on the Environment, organ of the Government of the State of Rio de Janeiro, FEEMA [6], performs the systematic monitoring once a fortnight in the Rodrigo de Freitas Lagoon. It is carried out in four sampling points for determination of some parameters, as, physical-chemical ones, biological ones (phytoplankton qualitative / quantitative) and the collection of sediment annually. Twice a week, some measurements are performed along the water column through vertical profiles in order to get the dissolved oxygen, salinity and temperature, to observe the mixture conditions of water and Secchi transparency. It is emphasized that this monitoring can be enhanced in the light of events observed mainly during the summer. Weekly inspections are also carried out on galleries of rainwater that flow into the lagoon and in the channel that links it to the sea.

The analysis of that environmental data series, particularly the events of fish death, enabled the development of a warning system for accidents with the ichthyofauna. This alert system is based on only four parameters: Dissolved Oxygen, Turbidity, water Temperature and Phytoplankton concentration.

Table 1: Monitoring parameters.

Indicators	Values		Risk index(RI)
OD surface	>5mg/l		0
	≤5mg/l		5
OD bottom	≤1mg/l	<3 layers	0
	≤1mg/l	≥3 layers	2
Secchi disk	>50 cm		0
	≤50 cm		1
Temperature	<30°C		0
	≥30°C		1
Phytoplankton	<70%	Dominance	0
	≥70%	Dominance	1
			∑ (10)

Table 2: Alert levels.

Categories	Reference ranges
Vigilance	RI < 2
Alert 1	2≤RI<4
Alert 2	4≤RI<6
Alert 3	RI≥6

3 Proposed work

The water quality is represented by a set of characteristics, usually measurable, of chemical, physical and biological nature. The aquatic ecosystems incorporate, over time, substances derived from natural causes, without any human contribution, where high concentrations are rarely found, which, however, can affect the chemical behavior of water and its relevant uses. Meanwhile, other substances discharged in water bodies by the anthropic action, as a result of the nearby occupation and the bad use of the soil may result in serious problems to the water quality, which requires research and investment for its recovery. Relations between these characteristics are object of study in various research areas, aiming to integrate different types of knowledge. In data mining literature, there are various surveys conducted in various places worldwide.

This study proposes to investigate the interrelations of different variables presented in the set of polluting substances in the Rodrigo de Freitas Lagoon. Its

final goal is to reach a methodology for the classification of water quality, which could be used in other water bodies. At the beginning of our study, we combined the data, which are derived from three different databases, measured in different days and frequencies. After that, the investigative process is initiated, where different techniques of data mining are used to group and classify the available data. According to [3], there is not a single algorithm to produce the best performance for all the problems. Therefore, it is important to investigate the best option for the problem in question.

4 Methodology

4.1 Data pre-processing

When starting a work involving knowledge discovery, we have to bear in mind that any process of data preparation must be implemented until we have the best database for our goal. Data with specific characteristics, with measurements performed with different frequencies, or on different days, force us to try to get the best use of the available dataset. Since we are starting a process of research, we need to have some knowledge on the incomplete or inconsistent data. Our biggest problem with respect to the quality of the data was the fact that this database presents a high percentage of missing data. Those data had this type of problem, because in certain times, the collection of the material to be examined, or some in situ measurements, could not be made. Operational problems, often aggravated by the lack of resources, were the main causes for the occurrence of missing values. Since we are at the beginning of the investigation and wanted a preliminary and quick response, it was used only those records matching the same date. The others were ignored. In a more advanced stage a specific function of interpolation will be employed in order to consider all the data.

Our database is composed of information from three sources: physical data, such as water temperature, turbidity, dissolved oxygen and salinity; chemical data, or levels of nutrients such as nitrate, nitrite and phosphorus, and biological data, consisting of concentration values of microorganisms.

After verifying distance matrices and distribution graphics, we have noticed that there isn't a strong correlation between the attributes of the database. All calculated values for correlation were close to zero. This was a surprising fact since our knowledge is that, for example, the presence of nutrients combined with high water temperatures, provides the appearance of phytoplankton. This is a known fact, but it could not be proven by the algorithms employed, which leads us to suspect that some premises we are following may not be the most appropriate.

In order to try to reduce the number of attributes, the Principal Component Analysis was performed. The results showed our previous knowledge that it was difficult to reduce the number of attributes. There isn't, in this case, a variable, or a group of few variables, which can represent statistically the information in the database.

5 Clustering

At the beginning of our investigation, a simple clustering algorithm was employed, which could give us a quick answer and an initial assessment of the problem we would be facing. So we decided by the k-means algorithm. Testing various options processing, we concluded that the best number of clusters for our database was three (3). This result was also achieved by the construction of a dendrogram and the calculation of the Calinski-Harabasz index [7].

According to the graphs of distribution, the attribute temperature seems to have been one of the most important attributes of the process of clusterization. Clearly we can see in the graph in Figure 2, that the cluster 2 presents a concentration in the highest values of temperature, giving maximum concentration around the temperature of 29°C and reaching up close to 33°C. In cluster 3, the concentration occurs in milder temperatures, around 24°C. The cluster 1 presents a set of values similar to the cluster 3, but without a pattern of concentration defined. The water temperature is a factor of great influence on the water quality, as it affects the concentration of microorganisms.

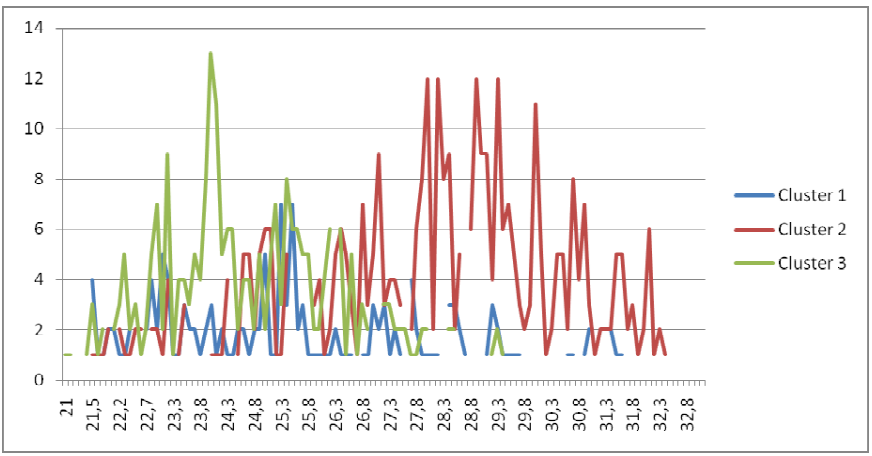

Figure 2: Temperature distribution.

In relation to the nutrients concentration, they took a decisive participation in the definition of cluster 1, especially the variable ammonia nitrogen, which presented the mean value twice the values of the other clusters. Turbidity also presented a significant weight in the definition of cluster 3, with mean value equal to 0.85 m, while the mean values of the other clusters presented values around 0.65 m.

An interesting point in this cluster analysis is the contribution of the Dominance attribute. This variable indicates if there is more than 70% of dominance in a type of microorganism over the others (1), or not (0). In an ecosystem, the different species of microorganisms must be well balanced. If there is a sort of dominance of one type over the other ones, this imbalance

shows that something is wrong. In this ecosystem, the dominance of cyanobacteria over the other microorganisms is the most frequent event. The cluster 2 shows an average of 0.10, which indicates that for this cluster, the dominance was small. The other two clusters did not present a similar behavior, with averages around 1.0, indicating a strong imbalance between the species.

Table 3: Mean values.

Salinity ‰	Oxygen Surface mg/l	Oxygen Bottom mg/L	Temperature ◦C	Turbidity Secchi disk	Phosp mg/L	Nitrate mg N/L	Nitrite mg N/L	Ammonia Nitrogen mg N/L
14.48	11.00	2.83	25.83	0.63	0.19	0.04	0.006	0.13
16.11	8.83	2.16	28.22	0.68	0.11	0.02	0.003	0.05
18.16	10.03	1.69	24.89	0.85	0.10	0.04	0.005	0.07

Nitrog Kjeldahl mg N/L	Ortho-phos mg/L	Solids Susp mg/L	Precipt accum (4 days) mm	Dominance	Cluster
1.84	0.02	38.04	19.33	0.75	1
1.22	0.02	27.90	18.64	0.10	2
0.94	0.01	41.88	10.20	0.81	3

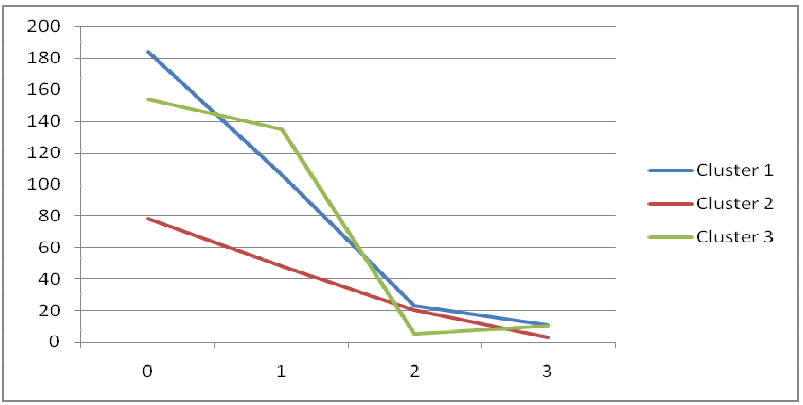

Figure 3: Actual classification X clustering analysis.

Figure 3 shows a framework for distribution of the current classification used by Feema and the clustering analysis of this work. In this Figure, the horizontal axis shows the current classification and the vertical axis, the number of occurrences for each cluster. Zero value in the horizontal axis indicates the status "vigilance" and values equal to 1, 2 and 3 indicate the increasing levels of "alerts". It should be noted that the actual criteria are very different from our results. In fact, the current methodology takes into account only a few variables and assigns different weights to them. In our clustering analysis it was employed a normalization of the variables values, and they were treated equally with the same weights. Moreover, the number of alerts (classes) currently employed is

equal to four while our results showed that the number of classes is only three. Indeed, we observed that alerts 2 and 3 could be clustered in just one class, since we evaluated the data with more variables than that actually employed.

6 Conclusions

Data on monitoring the water quality of the Rodrigo de Freitas Lagoon is collected in view of the constant concern about the quality and recovery of this lagoon complex. This work, so far, aimed to have an idea of the behavior of each indicator, its relations with other indicators and make an initial experience with clustering algorithms, with the purpose of discovering a more appropriate classification system for the water quality indices. At this stage, we met some difficulties with the available database that have not been solved at all.

As reported previously, the data are not displayed on a homogeneous way. There are groups of data that is collected on different days, with different frequencies. This characteristic limited us as some data are presented at a frequency so low that prevented us from using them along with the others.

Another difficulty encountered was the lack of information on certain external factors that can interfere decisively in the system, such as wind data and maneuvers in the locks on existing channels. These incidents may have an influence on the collected data and consequently in the results of our analyses.

As we have seen, the clustering algorithm separated our database into three clusters. It can be noted, reviewing the point's mean of each cluster, that cluster 2 has a trend to more favorable points in the balance of the microorganisms colony, with mean values of dissolved oxygen at surface lower than the other clusters, higher values of temperature and lower levels of dominance. Cluster 3 presented a reasonable rate of dissolved oxygen, milder water temperature and a high rate of dominance, besides a lower concentration of nutrients. Finally, cluster 1 had some average values next to the ones of cluster 3, but differs from it by showing highest levels of nutrients.

Analyzing the current classification system, we can say that Dissolved Oxygen attribute has a significant weight in it. This fact can help to predict the death of a great amount of fish by the lack of oxygen in the water, as it was usual a few years ago, but it can cover other types of problems, like an increase of the microorganism colony. Taking into account the dominance levels of microorganism, we can say that the algorithm has achieved the goal of separating the instances in groups according to a pattern of pollution, but there are other factors that should be examined more closely as the influence of the nutrients in the balance of the microorganisms' colony.

Some events have consequences that occur some time later. It is widely known that the presence of nutrients in the water such as phosphorus and nitrate, coupled with the low level of dissolved oxygen and high water temperatures, have strong influence in the development of colonies of photosynthetic organisms, such as cyanobacteria. This influence takes a while to appear, turning it difficult to be detected by the strategy employed here. Another time dependent event is the rain incidence, responsible for the input of organic material to the

water body (when the system rainwater combined with links of clandestine sewage is overloaded.), which should have influenced the results obtained. The rain, as well as the proliferation of microorganisms, has no immediate effect on the system turning it more complex to be detected by algorithms. In this preliminary study, the employed strategy may not be the ideal to handle this kind of problem, based on data collected in the same moment for each instance of the database. An algorithm that takes into account the temporal factor, which can predict trends in a mass of data, seems more appropriate. This point is part of our future works, but before doing that, we have to continue the currently investigation with a better and more detailed treatment in the existing data, so we can use all the information available, as well more significative classification rules.

References

[1] Robertson, DM, Saad DA, and Heisey DM. 2006. “A regional classification scheme for estimating reference water quality in streams using land-use-adjusted spatial regression-tree analysis.” Environmental Management 37 (2): 209–229.
[2] Rezende, Solange Oliveira; Intelligent Systems – Fundamentals and Applications; RECOPE-IA; Manole Ed.; 2003 (In Portuguese)
[3] Stark, J.R., Hanson, P.E., Goldstein, R.M., Fallon, J.D., Fong, A.L., Lee, K.E., Kroening, S.E., and Andrews, W.J. 2000. “Water Quality in the Upper Mississippi River Basin, Minnesota, Wisconsin, South Dakota, Iowa, and North Dakota, 1995–98.” United States Geological Survey, Circular 1211. http://pubs.usgs.gov/circ/circ1211/pdf/circular1211.pdf (last accessed July 4, 2006).
[4] Scheffer, M., S. Carpenter, J.A. Foley, C. Folke, B. Walker. 2001. “Catastrophic shifts in ecosystems.” Nature 413:591–596.
[5] Haykin, Simon. “Neural Networks – Principles and Practice” 2ª edition, Bookman; 2001
[6] www.feema.rj.gov.br
[7] Calinski, T., and Harabasz, J. (1974). "A Dendrite Method for Cluster Analysis", Communications in Statistics, 3, 1–27