Fixing the most representative physicochemical and biological variables of Medellín's Parque Norte's lake, as a basis for a mathematical modeling

Esteban Higuita-García
ehiguita@eafit.edu.co

Department of Mathematical Sciences
EAFIT University

***Abstract*—The Parque Norte's lake's water is supplied by polluted affluents, therefore use this lake is not allowed for primary or secondary contact activities with people, this pollution also affects animals who inhabit the lake's enviroment. This work has for principal objective provide statistical bases of main variables physicochemical that allow understand the pollution of lake for future work in bioremediation through a mathematical model, in order to recover the lake for recreational and sports activities. To accomplish this goal, the main parameters of water lake are identified through principal components analysis (PCA), which sets relations between the parameters and will reduce the dimensionality of the problem; this secondary information has provided by Metroparques, company who manages the park where the lake is located.**

***Index Terms*—Biological system modeling, Principal component analysis (PCA), lakes, Mathematical analysis**

I. Introduction

Most cities of the world have serious problems in air quality, which received increasing attention in recent years (Mayer 1999). However, cities also have other aspects that are severely contaminated, such as rivers and lakes present in around cities. In Medellín, Colombia, pollution levels are increasing in recent years regarding total suspended particles (TSP), whose levels exceed $100\mu g/m^3$, well above precautionary levels for health defined by the World Health Organization Health - WHO ($35\mu g/m^3$).

Lakes belong to calm and slow moving water category. By definition, a lake is a water extension surrounded by land everywhere (Roldán & Ramírez 2008). From limnological point of view, the lake being described like satin tropical system with eutrophic and hypertrophic features, which indicate this waters presents high nutrition concentration (Ramírez 1987). This characteristic is attributed as much the conditions of system like water quality situation deteriorating, generated for diffuse pollution sources discharging in the lake through surface and underground influx.

Pollution affects water quality in lakes and other freshwater resources around the globe. It can take many forms from industrial, agricultural, or municipal sources; a few common examples include pesticides, herbicides, sewage, and litter (Hicks 2014). For its determination is necessary define the water quality of the lake and determine the risks in human health, the standard requires that secondary contact for the number of total coliforms is $5000MPN/100ml$ fecal coliform (E. coli thermotolerant and E. faecalis) $1000cfu/100ml$, because its high affinity for gastrointestinal illnesses, high resistance to disinfection and responsible of causing infection in the urinary track and cutaneous pustules, also required that water not present films of any visible fat, oils and floating material resulting from human activities, toxic or irritating substances, whose act by contact, ingestion or inhalation produces adverse effects on human health (Metroparques & Secretaria de Medio Ambiente 2010).

Currently environmental management can not limited to control impacts, and must to consider the ecosystem in all its complexity, where many variables interact to produce changes that can not be predicted by analyzing only a part of itself. In this sense, environmental management can deepen their scope and consider the potential of restoring a ecosystem, which can be increased if it consider the new tools available to the researcher, such as bioremediation and ecotechnology (Jorgensen et al. 2005, chap. 5).

Parque Norte's lake is located in Medellín (Antioquia, Colombia) city, at coordinates $(06^o17'N, 75^o33.4'W)$ (Ramírez & Abel 1995). It is located within the park of the same name and has become an urban reference among the inhabitants of the Aburrá Valley as ecological, landscape and recreational scenario, is considered one of the largest urban lakes in Latin America (*Parque Norte* 2015). Despite this, the lake is still highly contaminated making it difficult to use the lake for sports and recreational activities; the lake has problems such as water hardness, high concentrations of iron that it confers considerable smell and taste, besides presence of compounds that could be toxic as arsenic (Metroparques & Secretaria de Medio Ambiente 2010).

In the study conducted by Antioquia University to Metroparques (Metroparques & Secretaria de Medio Ambiente 2010), inflows of Parque Norte's lake were characterized and found the system is fed by underground flows from the south-southeast, by the aquifer associated with the Aburrá river, that runs on the western side of the lake and to a lesser extent by groundwater flow from the east of the city (Metroparques & Secretaria de Medio Ambiente 2010). The same diagnosis recognizes that it not quantified the contributions of each incoming to the pollutants concentration in lake flows, neither has done a study to measure system response to changes in environmental conditions generated by the seasons of rain and drought, among others. Also is not available a supervising tool that provides insights into strategies to followed in the management of lake, and there is not the ability to predict the effectiveness of an intervention in it. This is a severe limitation if it want to establish a program of quality improvement with strong and enduring goals in time. The pollution of groundwater is caused by the presence of urban waste water who are mainly fecal, then often the presence of E. coli and Pseudomonas aeruginosa, along with nitrates, represent the increasing content of total salts and chlorides. The presence of nitrates usually occurs when wastewaters discharged directly into the unsaturated zone and creates an oxidizing environment. If the conditions are anaerobic or direct discharge on the aquifer, is produced decrease of nitrate content with the presence of $Fe2+$ dissolved(Foster & Hirata 1988).This is especially important if it take the presence of the Aburrá river near the study area, this is due Aburrá river receives discharge of six municipalities of the Metropolitan area, before going through to park and surrounding aquifer, additionally the site of the lake itself been affected by former litter disposal sites (Metroparques & Secretaria de Medio Ambiente 2010).

Under this issue, Metroparques have raised several recent studies to Parque Norte's lake, from the temporal variation of the trophic status (Moreno & Ramírez 2010) since the temporal variation of the trophic status, to interventions to implement environmental strategies that contribute to sustainable management of lake (Metroparques & Secretaria de Medio Ambiente 2010). It can not find in literature studies from a mathematical view, such as implementing of a mathematical model as a understanding tool, even though in other tropical lakes has been used as referent to study their behavior (Parinet et al. 2004).

Mathematical models are increasingly used in environmental management, mainly because models allow to link quantitatively the impact exerted on a ecosystem with the changes in ecological status of it. In particular, the models of aquatic ecosystems had a boom in recent decades and have found a wide range of applications in the management of lakes and reservoirs (Jorgensen et al. 2005, chap. 5). The calibration of big complex models depends on assumptions and arbitrary decisions, and often the calibration procedures do not adequately address the uncertainty in data that describe the study under the system (Kurt et al. 1981). For a correct construction of mathematical model, it should conduct an analysis of main variables that describe it and this is where this research focuses (Parinet et al. 2004). This information together with data likelihood analyzed, are close to reality, therefore much of uncertainty and arbitrariness is inevitable on ecological and theoretical modeling limitations; these procedures must include dependence between the variables included in predictions and sensitivity to uncertainty and arbitrary assumptions (Somlyódy 1982).

Metroparques, company responsible for over 25 years Parque Norte's lake (Metroparques & Secretaria de Medio Ambiente 2010) in its search solve the problem of pollution of the lake, it hires to In Sithu company that is abiding by this premise, it intending the mathematical modeling to be established as tool diagnosis and management of this ecosystem; trying to visualize the causes and mechanisms that drive the deterioration of the ecological quality of the system. This knowledge will allow evaluating and selecting intervention strategies that minimize or reverse the negative impacts on the system, intervening as little as possible in its natural functioning and looking for process economics; keeping this in mind will give priority to using bioremediation techniques that enhance natural self-purifying capacity of the system, as well as certain economic criteria and sustainability of the process over time, the model also can be used to assess ecosystem functioning under these or other conditions of intervention, which may be taken up in future studies.

II. Methodology

The modeling process begins with a review of secondary information that addresses the issue of the aquatic component of the lake, such as: water balance, limnological characterization, health status and environmental strategies that contribute to the sustainable management of Parque Norte's lake located in Medellín city, among others. Through the review and analysis of information from investigations in this aquatic ecosystem, quantitative data is selected to be incorporated in a series of matrices for further exploratory analysis of statistical type. The methods used to collect the information, the frequency of sampling, experimental design, the variability and significance of their results are evaluated.

Table 1 contains the abbreviations of the terms that were used in the article

The data analyzed, table I come from the investigation conducted by Professor Jhon Jairo Ramirez between March and August 1986 where surface water samples were taken to measure the physical, chemical and biological parameters in the Parque Norte's lake in one station near from the center.

Table II contains the abbreviations of the terms that were used in the article.

Term	Abbreviation
Transparency	Tr
Conductivity	Cd
Water temperature	wT
pH	pH
Dissolved Oxygen	DO
Nitrogen as nitrate	Nn
Total Phosphate	tP
Total hardness	tH
Chlorides	chl

Table II – *Parameters data*

Once structured suitable matrices, multivariate statistical methods were applied to a large percentage, including Principal Component Analysis (PCA), factor analysis (FA), Canonical Correspondence Analysis (PCA), Cluster Analysis (CA)

parameter/ sample	**Tr**	**Cd**	**wT**	**pH**	**DO**	**Nn**	**tP**	**tH**	**Chl**
1	0.22	2.6e+03	22	8.8	1.2	0.05	0.37	1.2e+02	0
2	0.2	2.6e+03	22	8.7	1.2	0.02	0.45	1.3e+02	0
3	0.4	2.4e+03	23	8.7	2.6	0.05	0.47	1.3e+02	0
4	0.4	2.4e+03	24	8.3	1	0.12	0.2	1.4e+02	6.2e+02
5	0.35	2.4e+03	24	7.9	4	0.1	0.2	1.3e+02	5.8e+02
6	0.44	2.4e+03	25	8	3.6	0.09	0.3	1.1e+02	5.9e+02
7	0.3	2.4e+03	24	7.9	3.6	0.06	0.3	1.2e+02	6.1e+02
8	0.3	2.5e+03	24	8.5	2.9	0.02	0.8	1.1e+02	7.6e+02
9	0.25	2.5e+03	26	6.9	0.8	0.05	0	1.2e+02	5.5e+02

Table I – *Parameters data*

and Canonical Correlation Analysis(ACCA) (Rencher 2003). The purpose of this process lies in finding the most representative and key variables in the ecological functioning of the lake, and understand the relationship between them. In this research only the PCA is made. The resulting statistical analysis, as mentioned above, is the first step to structure the dynamic model that modulates the ecological functioning of the Parque Norte's lake; that is, with key variables and variables with higher weight, an iterative process is carried out with different mathematical models existing in the literature (Jorgensen et al. 2005) , aimed to determine the ecological quality of water, trying to take a glimpse on the research problem that revolves around pollution of water and sediment matrix. For the next phase, the experimental design of higher affinity will be configured to measure and account for variables that are modulating matrix contamination of water and sediments in a lake basin, which come to form and generate a model of a dynamic system of nonlinear differential equations. To start the analysis of data collected in the literature a standardization process is performed and arranged in column and row parameter test it sets out to observe the distribution of measurement data. You must check the correlation of couples between the variables, if any variables with high correlation, that is, over 80% correlation may be performed PCA which build new independent variables that are linear combinations of the original variables . To perform this analysis is appropriate to calculate the main component raw data when all the variables are in the same unit. Because when the variables are in different units or the difference in the variance of different columns is substantial (as here), the weighting of the data or using the weights is preferred. The PCA is performed using the inverse variance parameters as weights, using equation 1.

$$w = \frac{1}{VAR(parameters)} \tag{1}$$

We proceed to apply the PCA which results in the matrix of principal components, the score for each component variability, the confidence level that explains each vector and a statistical measure of the multivariate distance of each observation from the center of the whole database called Hotelling's T^2. Finally we analyze each of these results from a statistical point of view and limnology.

III. Results

In the first step it was found that the vast majority of data is highly concentrated as shown in the figure 1.

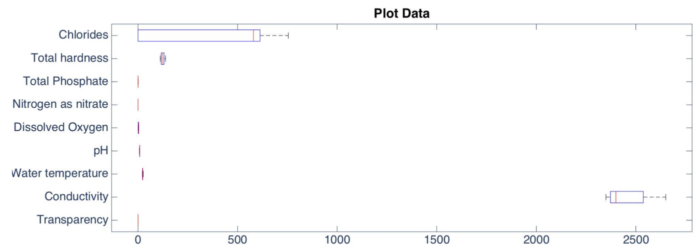

Figure 1 – *Box Plot all parameters*

However when an individual analysis does not identify outliers, in the figures 2, 3and 4.

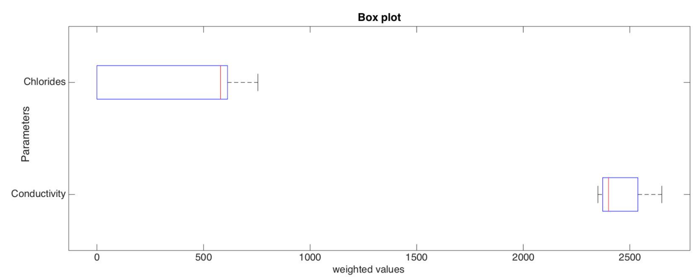

Figure 2 – *Box plot some parameters*

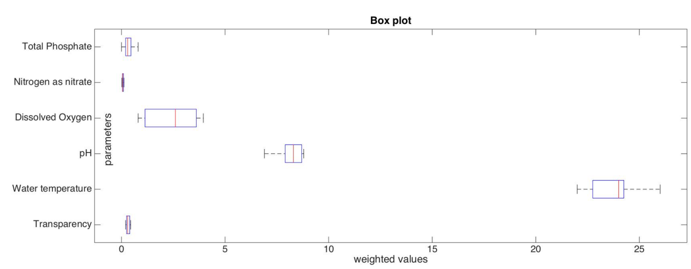

Figure 3 – *Box plot some parameters*

Chlorides and conductivity, which particularly exhibit a high dispersion, are highly related in chemical because chlorides are ions and their concentration indicates a greater capacity for electrical conductivity of water. In reviewing the correlation in the data array such Figure, the correlation between some variables is small, however, in other variables is as high as 82%. This allows to infer that an PCA can be implemented. Applying this analysis it is obtained as a result a component scores matrix 5.

Taking the first two vectors coefficients, it are weighted

1	-0.83	0.39	-0.05	0.51	0.68	-0.069	0.051	0.37
-0.83	1	-0.57	0.36	-0.61	-0.57	0.23	0.0068	-0.52
0.39	-0.57	1	-0.92	0.17	0.4	-0.58	-0.35	0.72
-0.05	0.36	-0.92	1	-0.047	-0.26	0.72	0.26	-0.58
0.51	-0.61	0.17	-0.047	1	0.2	0.22	-0.28	0.4
0.68	-0.57	0.4	-0.26	0.2	1	-0.59	0.44	0.39
-0.069	0.23	-0.58	0.72	0.22	-0.59	1	-0.3	-0.12
0.051	0.0068	-0.35	0.26	-0.28	0.44	-0.3	1	-0.32
0.37	-0.52	0.72	-0.58	0.4	0.39	-0.12	-0.32	1

Table III – *Component scores matrix*

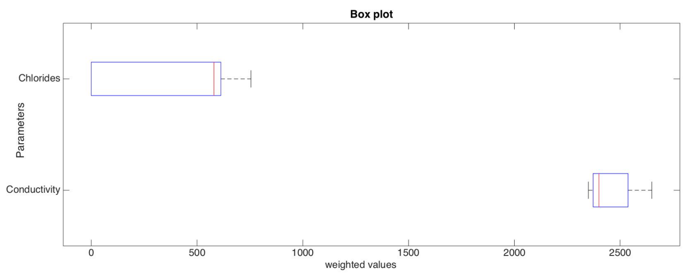

Figure 4 – *Box plot some parameters*

components, implying, the coefficient matrix is orthonormal, which are transformed in this way using equation 2

$$coefforth = diag(\sqrt{w})) \times wcoeff \quad (2)$$

The score contains the coordinates of the original data in the new coordinate system defined by the principal components. Score matrix is the same size as the input data matrix. Since these scores are obtained using components orthonormal coefficients and standardized ratings using the equation 3.

$$cscores = Standardized(parameters) \times coefforth \quad (3)$$

The first two principal components alone are evaluated. PCA calculate the scores have mean zero. By plotting these components must identify the points of the rightmost of plot 5. It is noted that the tests 4,5,6,7 and 9 are the best allows identify the major components obtained.

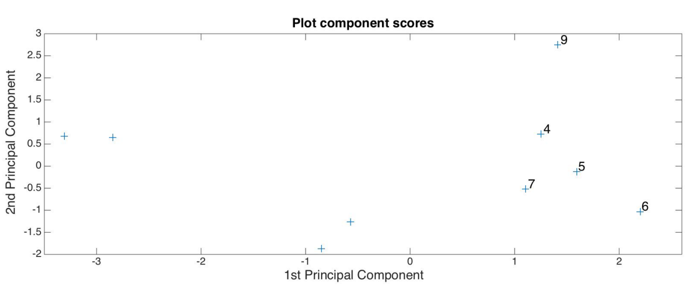

Figure 5 – *Plot component scores*

However it evaluate the variance explained by the corresponding principal component. Each column of score has a sample variance equal to the corresponding row of latent, in table IV.

4
1.9
1.8
0.48
0.42
0.26
0.058
0.0018

Table IV – *Latent vector*

45
21
20
5.3
4.7
2.8
0.65
0.019

Table V – *Explained vector*

And the explained variance (in table V) is a vector containing the percent variance explained by the corresponding principal component. These indicate that you can choose the first two components because the variance explained by obtaining three components results in a 86.53% which exceeds 80% of recommended data collection. In the table VI are the fist two principal components.

1	-0.83
-0.83	1
0.39	-0.57
-0.05	0.36
0.51	-0.61
0.68	-0.57
-0.069	0.23
0.051	0.0068
0.37	-0.52

Table VI – *First two principal component matrix*

This scree plot 6 only shows the first five (instead of the total eight) components that explain 95% of the total variance. The only clear break in the amount of variance accounted for by each component is between the first and second components. However, the first component by itself explains less than 5% of the variance, so more components might be needed. You can see that the first two principal components explain roughly two-thirds of the total variability in the standardized ratings, so that might be a reasonable way to reduce the dimensions.

The first component comprises as Water temperature, Nitrogen as nitrates, Chlorides, Transparency, Dissolved Oxygen. In turn the second component is composed Chlorides,Transparency, Dissolved Oxygen,Total Phosphate, pH. Conductivity and Total hardness parameters are not well explained by either components, in change Chlorides,Transparency and Dissolved Oxygen may contribute to

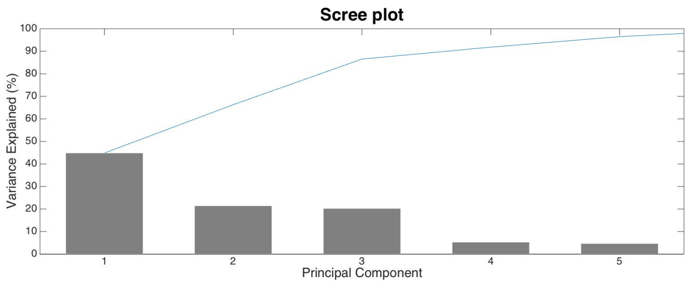

Figure 6 – *Scree Plot*

the two components. See more at 7.

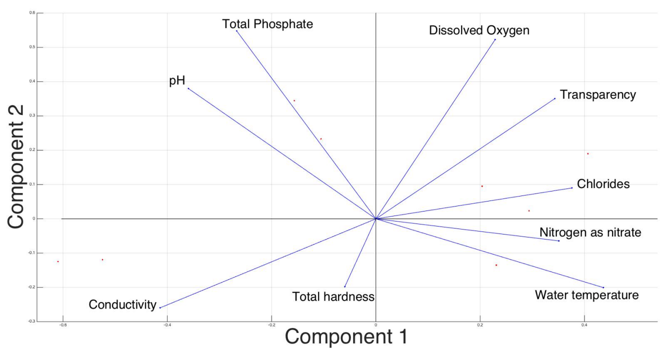

Figure 7 – *Principal components*

In order to find the most extreme data points is applied Hotelling's T^2, and is obtained Transparency it is.

IV. Conclusion

- Hypotheses are proposed to characterize the resulting components:
  The first component indicates the productive capacity of the system and it availability of dissolved oxygen by which the nitrogen is mineralized and reaches nitrates forms, reason why was named as Photosynthetic rate component.
  The second component indicates the compounds concentration that are mineralized; considering chlorides have a very low contribution, this component is named as water mineralization.
- From the results are open questions to resolve issues that resources in this investigation are not resolved:
  How chlorides behave with respect to high temperature, dissolved oxygen and nitrates.
  It can be classified Parque Norte's lake as a wetland in their limnological atypical behavior compared to other water bodies in the sector.
  The relationship of pH and total phosphates indicate good nutrient availability.

Acknowledgments

I want to thank Dr. Nelson Gil for their trust and support in the preparation of this work, the engineer Wendis Torreblanca its availability for data collection and secondary data, Professor Julio Hurtado for the guidance provided from the biological perspective and Professor Francisco Zuluaga statistics to provide guidelines for proper analysis.

Anenexes

The following link is a folder with references, sketches, graphs obtained and code in Matlab R2014b: https://goo.gl/uQjfU7

References

Foster, S. & Hirata, R. (1988), *Determinación del riesgo de contaminación de aguas subterráneas: una metodología basada en datos existentes*, OPS.
URL: *http://goo.gl/ZxO6zj*

Hicks, S. (2014), 'Water quality: Pollution', http://www.lakescientist.com/lake-facts/water-quality/. [Accessed: 15th February 2015].

Jorgensen, S., Loffler, H., Rast, W. & Straskraba, M. (2005), *Lake and reservoir management*, Vol. 54, Elsevier, chapter The use of mathematical modelling in lake and reservoir management.

Kurt, F., Gerrit, V. & Beck, M. (1981), 'Uncertainty and arbitrariness in ecosystems modelling: A lake modelling example', *Ecological Modelling* **13**(1–2), 87 – 110.
URL: *http://goo.gl/26Dh9o*

Mayer, H. (1999), 'Air pollution in cities', *Atmospheric environment* **33**(24), 4029–4037.

Metroparques & Secretaria de Medio Ambiente (2010), Convenio interadministrativo de cooperacion para la implementacion de estrategias ambientales que aporten al manejo sostenible del lago del parque norte localizado en el municipio de medellin.

Moreno, D. & Ramírez, J. (2010), 'Temporal variation of trophic condition in lake of the parque norte, medellín (antioquia),colombia', *Actualidades Biológicas* **32**(92), 75–87.

Parinet, B., Lhote, A. & Legube, B. (2004), 'Principal component analysis: an appropriate tool for water quality evaluation and management—application to a tropical lake system', *Ecological Modelling* **178**(3–4), 295 – 311.
URL: *http://goo.gl/1lgAnK*

Parque Norte (2015), http://www.parquenorte.gov.co/web/. [Accessed: 13th March 2015].

Ramírez, J. (1987), 'Contribución al conocimiento de las condiciones limnológicas de la laguna del parque norte', *Actualidades Biológicas* **16**(59), 12–27.

Ramírez, J. & Abel, D. (1995), 'Cambios diurnos de temperatura y variables físicas y químicas en dos épocas delaño en la laguna del parque norte, colombia', *Acta limnologica brosìlìtnsia voi vii ê7˜ i04* p. 87.

Rencher, A. (2003), *Methods of multivariate analysis*, Vol. 492, John Wiley & Sons.

Roldán, G. & Ramírez, J. (2008), 'Fundamentos de limnología neotropical', *Colección Ciencia y Tecnología. Editorial Universidad de Antioquia. Colombia* pp. 77–94.

Somlyódy, L. (1982), 'Modelling a complex environmental system: The lake balaton study', *Mathematical Modelling* **3**(5), 481–502.