ASSESSMENT OF SURFACE WATER QUALITY USING MULTIVARIATE STATISTICAL TECHNIQUES: A CASE STUDY OF SAIGON RIVER

Nguyen Huu Quang, Vuong Quoc Phuong, Pham Thi Ngoc Anh, Tran Thi Thuy Tien,
Ho Thi Phuoc and Truong Lam Son Hai*
Faculty of Chemistry – University of Science HCMC

[*] Truong Lam Son Hai, email: tlshai@hcmus.edu.vn

ABSTRACT

Analysis and management of surface water quality is a need for many economic and production fields, but requires much time and forces. Multivariate statistical algorithms are applied to the dataset, which made up from 19 water quality criteria collected from 10 sampling sites across waterways from Sai Gon river basin. PCA-X (PCA – **P**rinciple **C**omponent **A**nalysis) model of the dataset provides grouping by geographical location and flow direction, with explanation of the first 2 principal components are 62.4 and 25.2 %, respectively, which overviews the quality of water of these sampling sites, and allows determination of unexpected pollution sources from the system. These results are the basis of developing a method for delimiting and securing local pollution sites, assisting water quality monitoring and environmental management.

Keywords: multivariate, water quality, environment, pollution control, Sai Gon river

TÓM TẮT

Phân tích và quản lý chất lượng nguồn nước mặt là công việc cần thiết phục vụ cho nhiều ngành kinh tế và sản xuất, đòi hỏi nhiều thời gian và nhân lực. Các thuật toán xử lý dữ liệu đa biến được áp dụng với bộ dữ liệu thu thập từ các phương pháp phân tích xác định khác nhau với 19 chỉ tiêu chất lượng nguồn nước tại 10 điểm lấy mẫu trên hệ thống kênh rạch thuộc sông Sài Gòn. Mô hình PCA (phân tích thành phần chính) của bộ dữ liệu trên cho kết quả phân nhóm theo vị trí địa lý thành công với mức độ giải thích qua hai thành phần chính lần lượt là 62.4 % và 25.2 %, khái quát được toàn cảnh chất lượng nước tại các điểm lấy mẫu và cho phép xác định các vị trí có mức độ ô nhiễm bất thường. Những kết quả trên là tiền đề cho việc xây dựng một phương pháp nhằm khoanh vùng, định hướng phát hiện nguồn ô nhiễm cục bộ, hỗ trợ cho công tác quản lý môi trường và giám sát chất lượng nguồn nước.

Từ khoá: dữ liệu đa biến, chất lượng nước, môi trường, quản lý ô nhiễm, sông Sài Gòn

INTRODUCTION

Sai Gon - Dong Nai river system, which is the most important drinking water of Ho Chi Minh City (HCMC) and nearby areas, are severely polluted. According to HCMC Environmental Protection Agency (HEPA), monitoring results showed that the quality of water supply facilities of the Saigon river system is severely impaired. The pollution of this river system has been reported for years. A recent survey in 22 districts showed Saigon river water pollution has overpassed alarming levels. This river receives hundreds of thousands of cubic meters of unprocessed industrial, domestical and livestock wastewater every day. Meanwhile, according to the Saigon Water Supply Corporation (Sawaco), Saigon - Dong Nai river system provide raw water for domestic treatment plants of HCMC such as Tan Hiep, Thu Duc, BOO Thu Duc. Those are the source of drinking water for both the nearly 10-million people city, also nearby growing regions such as Binh Duong, Dong Nai province. Thus, the survey, analyze and evaluate pollution of the two river systems have been an urgent demand. To monitor the pollution of surface water, a multi-stage process had to be developed and applied, including: water sampling process, analysis of physical-chemical-biological criteria and statistical evaluation of the results. In that process, analysis and evaluation of physical-chemical-biological criteria are extremely important.

In order to identify local sources of pollutants and improve environmental quality around the waterway areas, the research team analyzed 19 physical-chemical-biological criteria of surface water samples, then apply multivariate data algorithms to evaluate the distribution of the sampling point's water quality. PCA model shows 2 unexpected polluting locations with abnormal levels, which propose a method for monitoring the aquatic environment results with speed, accuracy and stability.

MATERIALS & METHODS

Surface water was sampled from 10 specifically selected from 3 inland waterways system, including: Tan Hoa – Lo Gom (TH), Tau Hu – Ben Nghe (BN) and Nhieu Loc – Thi Nghe (NL). These are 3 of all 5 largest drainage system of the city with the total length of 55 kilometers, gathering more than 70 % of waste water from the urban and industrial districts (Minh 2009).

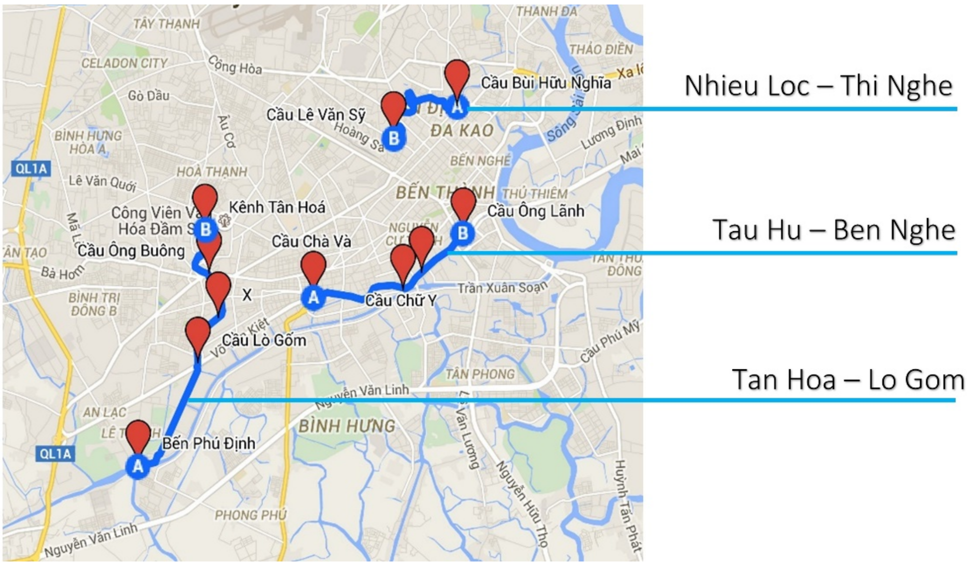

Figure 1: Waterway systems (dark blue) and sampling points (red)

Nhieu Loc – Thi Nghe system have the estimated length of 9.47 km, is the natural drainage system for many urban districts. The waterways are affected by the unusual semidiurnal tide mode of the South China Sea, and also by the minor difference of geographical height between the upstream and downstream. Slow flow rate often gets floating plants and garbage to stuck at some point where the stream curved, which produce pollution. The system also has some small streams, including Van Thanh, Bui Huu Nghia, Cau Bong and Ong Tieu canals.

Tau Hu – Kenh Doi – Kenh Te and Ben Nghe canals are two systems located in the south of the city's downtown. Total length of the two systems are 19.5 and 3.15 km, respectively. The widest cross-section of the canal is about 88 – 92 km, height from the bottom to the surface spread from 1.9 to 2.2 meters. The system is limited on both sides by Can Giuoc River and Sai Son River. Industrial and domestic wastewater from urban districts usually enter the stream without proper treatment. Due to tidal influences from both Can Giuoc and Sai Gon River, the area's hydrology is very complex, and water regime can change drastically during the day.

Tan Hoa – Lo Gom system have the mainstream of roughly 7.240 km with NE – SW flow direction, crossing Tan Binh District, District 11, 6, 8, and end jointing with Tau Hu downstream. The canal has a base height of 2.5 to 3.0 m, also affected by the semidiurnal waves. Flow rate during the dry seasons drop to below 10m3/s, and stay up to 30 m3/s during rainy season. 4 renowned industrial clusters are located by the mainstream, including 30% food processing and 28 % plastic & rubbery companies have direct wastewater discharge to the entire system.

All water samples were sampled following TCVN 5999:1995-2 (equivalent to ISO 5667-10:1992) standards. Samples for chemical analysis were acidified immediately, conditioned and storage at 4 ^oC. Ion chromatography samples were pre-filtered by 0.45 µm CA syringe filter and analyze in 24 hours. Water quality of the sampling sites are separated by 3 main sampling sessions conducted in December 2015, with 19 criteria are analyzed. The criteria are selected according to sample stability, properties of the drainage systems and conditions of the laboratory. 4 main analytical methods used for the analysis are: direct instrumental results (pH, conductivity, TDS, TSS, total hardness), spectrophotometry (NH_4^+, NO_2^-, F^-, Fe^{2+}, PO_4^{3-}, total Phosphorus), atomic absorption/emission spectrometry (Na^+, K^+, Zn^{2+}, Mn^{2+}) and ion chromatography (Cl^-, Br^-, NO_3^-, SO_4^{2-}). Official standard methods (Vietnam Standards, Standard Methods for Examination of Water and Wastewater 2002 - SMEWW, US EPA Methods) were applied with strict quality control by blank samples, matrix spikes and recovery experiments.

Raw water quality data are subjected to z-scale transformation according to (Shrestha & Kazama 2007; Alberto et al. 2001) to avoid clustering bias caused by significant differences of the mean values and deviations of the variables in the dataset. Statistical pretreatments techniques and Principal Component Analysis (PCA) were performed using Microsoft Excel 2013, Umetrics SIMCA-P 11 and IBM SPSS Statistics 23.

RESULTS & DISCUSSIONS

Goodness-of-fit of the data to normal and log-normal distributions were tested by Kolgomorov-Smirnov (K-S) statistical test. 13 of 19 variables collected are log-normal distributed with 95% confidence level. This result is the basis for choosing data transformation algorithms for each of the variable. Compatibility to PCA of the data were evaluated using Kaiser-Meyer-Olkin measure of sampling adequacy (KMO-MSA), with overall MSA = 0.734 suggests that the correlations between variables are not unique, and the data is well distributed in the PCA model. Bartlett test of sphericity provides Sig = 0.0000 (<< 0.05), indicate suitability of the data for structure detection.

Analysis data undergoes a preliminary assessment step using Spearman-R correlation matrix (Shrestha & Kazama 2007). 83 % of the variables are non-parametrically correlated with the confidence level of 70 %, of which 69 % have p-value > 0.95.

Table 1: Descriptive statistics results and analytical methods of the collected data

Criteria	Unit	Average	SD	LOD	Analytical method(s)
pH	--	7.17	0.09	--	Direct pH-meter readings
NO_2^-	mg $N.L^{-1}$	0.16	0.23	0.005	Spectrophotometric and Ion chromatography
NO_3^-	mg $N.L^{-1}$	62.67	29.68	1.50	Spectrophotometric and Ion chromatography

NH_4^+	mg N.L^{-1}	9.16	8.64	0.080	Spectrophotometric
Cl^-	mg.L^{-1}	479.4	345.4	--	Titrimetric
F^-	mg.L^{-1}	1.09	0.82	0.030	Spectrophotometric and Ion chromatography
Br^-	mg.L^{-1}	1.17	0.98	0.020	Ion chromatography
SO_4^{2-}	mg.L^{-1}	105.3	45	0.50	Ion chromatography
ortho-PO_4^{3-}	mg.L^{-1}	0.41	0.52	0.050	Spectrophotometric and Ion chromatography
Total Phosphorus	mg.L^{-1}	1.01	0.94	0.120	Spectrophotometric
Total Hardness	mg $CaCO_3$.L^{-1}	430.3	218.8	--	Titrimetric
TDS	mg.L^{-1}	937.8	547.1	--	Gravimetric
TSS	mg.L^{-1}	51.5	30	--	Gravimetric
Conductivity	µS.cm^{-1}	1780.1	997.8	--	Direct conductivity meter readings
Iron	mg.L^{-1}	1.16	0.69	0.050	Spectrophotometric
K^+	mg.L^{-1}	28.76	6.46	0.80	Flame AES
Na^+	mg.L^{-1}	278.3	164	0.16	Flame AES
Zn^{2+}	mg.L^{-1}	0.23	0.21	0.030	Flame AAS
Mn^{2+}	mg.L^{-1}	0.34	0.17	0.060	Flame AAS

*LOD: Limit of detection

Principal Component Analysis (PCA) is a multivariate data analysis method, on the basis of reducing the number of dimensions, by defining new variables consisting of linear combinations of the original ones, in such a way that the first axis is in the direction containing most of the variation (Wehrens 2011). PCA allows summary and overview of the distribution of the data, in order to deploy prediction models (Abdi & Williams 2010) and quality control (Saporta & Niang 2009).

Logarithm-transformed variables using KS index are imported into Umetrics SIMCA-P 11 software program, and apply a proper PCA-X method. 5 variables which is pointed possible outliers are excluded to increase R^2 and Q^2 of the models, including pH, NO_3^-, total hardness, TSS and Zn^{2+}. With all remain variables, a “scree plot” contains 6 of the first components are described in Figure 2.

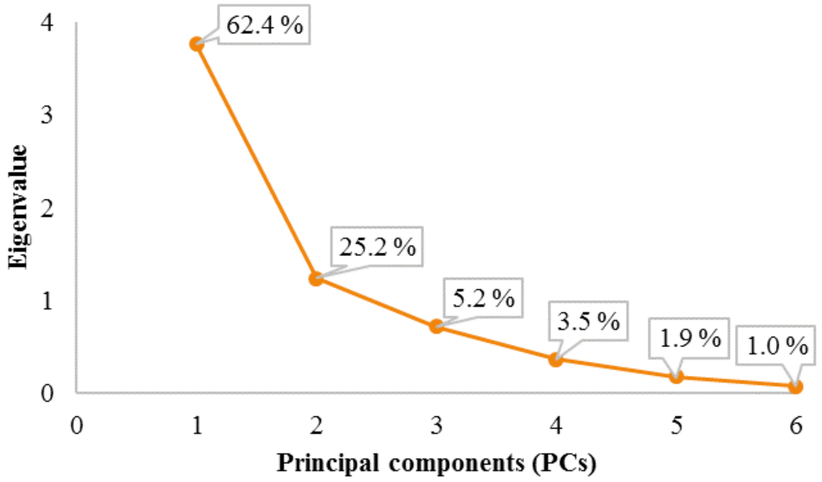

Figure 2: Scree Plot of the optimized data

Scree Plot of the model hints that 87.6 % (PC1: 62.4 %, PC2: 25.2 %) variances of the data are described by first 2 PCs, therefore the PCA model only requires PC1 and PC2 to be declared to summarize the distribution of the chemical-physical criteria. Principal component 3 have explanation of roughly 5 % and decreases with PC numbers, usually does not require further analysis (Arslan 2013; Christirani et al. 2015).

A 2-dimensional space include projections of the data points to PC1 and PC2 (often called "Score plot"), are shown on Figure 3. Clear evidence of clustering by waterways system and location of the samples is described, and data points are divided in 3 groups:

a. Nhieu Loc – Thi Nghe group (NL-TN): include 4/4 data points of the system, with slight dispersion and does not overlap to other groups

b. Tan Hoa – Lo Gom group (TH-LG): with a sub-group created by two sampling locations (Tan Hoa and Ong Buong), and an unexpected individual went in the center of the plot. This clearly indicates a severely polluted status of Lo Gom area, which is notorious in the recent years caused by lack of water treatment facility and highly populated discharge sources.

c. Tau Hu – Ben Nghe (BN): A sub-group is formed containing both the inner system points (NVC, CY) and the intersection point "Phu Dinh" between Tan Hoa and Ben Nghe system. Further study of the currents and hydrological properties of the two system show that the intersection point actually inherit some properties from the latter.

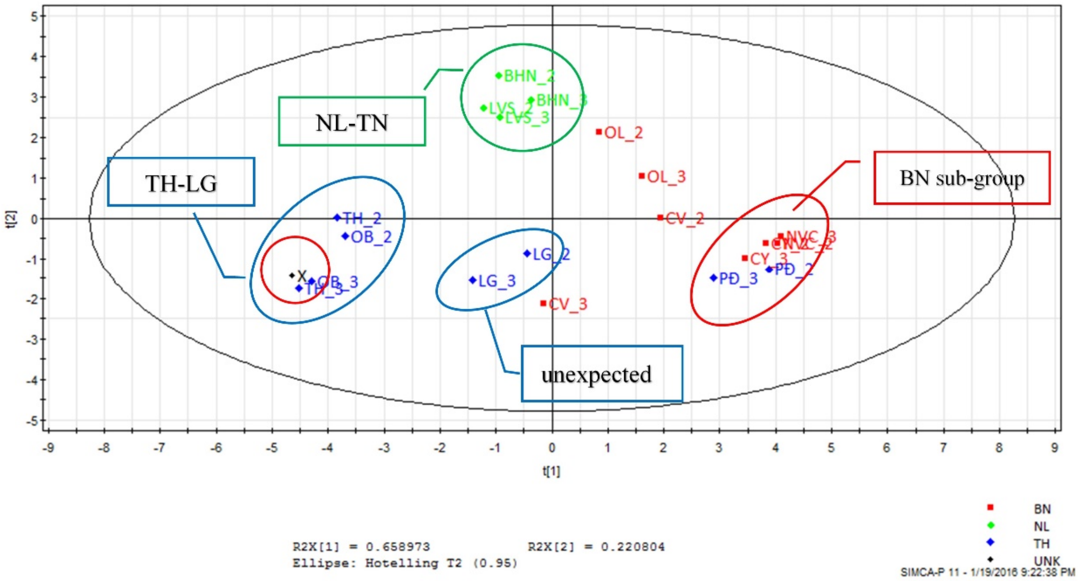

Figure 3: Score plot of PC1 (horizontal axis) versus PC2 (vertical axis)

In order to examine patterns recognition abilities of the model, an independence sample "X" is analyzed, which is located at the midpoint between Ong Buong and Lo Gom (Figure 1). Projection of "X" in the same model indicates that the water sample of X has the same properties with the water sample from Tan Hoa and Ong Buong. This result is a remarkable basis to develop a method to calculate the estimated position and narrow the search scope for unexpected pollution sources.

CONCLUSIONS

Results obtained from the application of multivariate statistical algorithms allow detection of 3 polluted positions in the main waterway systems. Stability and predictability tests proves the potential of this research scheme, suggests long-term applications for water quality management and law enforcement.

ACKNOWLEDGEMENTS: This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number C2015 – 18 – 10.

REFERENCES

Abdi, H. & Williams, L.J., 2010. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4), pp.433–459.

Alberto, W.D. et al., 2001. Pattern Recognition Techniques for the Evaluation of Spatial and Temporal Variations in Water Quality. A Case Study: : Suquía River Basin (Córdoba-Argentina). *Water Research*, 35(12), pp.2881–2894.

Arslan, H., 2013. Application of multivariate statistical techniques in the assessment of groundwater quality in seawater intrusion area in Bafra Plain, Turkey. *Environmental Monitoring and Assessment*, 185(3), pp.2439–2452.

Christirani, S., Zaharin, A. & Kamil, M., 2015. Classification of River Water Quality Using Multivariate Analysis. *Procedia Environmental Sciences*, 30, pp.79–84. Available at: http://dx.doi.org/10.1016/j.proenv.2015.10.014.

Minh, N.T., 2009. Hệ thống kênh rạch tiêu thoát nước của TP.HCM. *yeumoitruong.vn*. Available at: http://www.yeumoitruong.vn/threads/he-thong-kenh-rach-tieu-thoat-nuoc-cua-tp-hcm.6869/ [Accessed February 6, 2016].

Saporta, G. & Niang, N., 2009. Principal component analysis: application to statistical process control. *Data Analysis*, pp.1–23.

Shrestha, S. & Kazama, F., 2007. Assessment of surface water quality using multivariate statistical techniques: A case study of the Fuji river basin, Japan. *Environmental Modelling and Software*, 22(4), pp.464–475.

Wehrens, R., 2011. *Chemometrics with R: Multivariate Data Analysis in the Natural Sciences and Life Sciences*,