# APPLICATION OF MULTIVARIATE STATISTICAL ANALYSIS METHODS TO THE STUDY ON THE HUMAN IMPACTS ON THE WATER QUALITY OF THE DU TRIBUTARY IN THE CAU RIVER BASIN

ỨNG DỤNG PHÉP PHÂN TÍCH THÔNG KÊ ĐA BIẾN TRONG NGHIÊN CỨU ẢNH HƯỞNG CỦA CON NGƯỜI ĐẾN CHẤT LƯỢNG NƯỚC TRÊN SÔNG ĐU, THUỘC LƯU VỰC SÔNG CẦU

***Hoang Thu Huong[1], Koen Lock[2], Peter Goethals[2], Dang Kim Chi[1]***

*(1) Hanoi University of Technology; (2) Ghent University*

## ABSTRACT

*Monitoring program is crucial aspect of the river management. However it is more important to make interpretation of the monitoring data in order draw an appropriate management measures for restoration and improvement of river water quality. Within the context of this study, we posed two problems: (1) identification of the factors that influence the water quality of the river and (2) clustering monitoring sites according to ecological health. Resolving these two problems necessitated the application of multivariate statistical analysis methods, including parametric and non-parametric correlations, canonical correlation analysis and hierarchical cluster analysis using Bray-Curtis distance. These methods thus are being applied to analysis on human impacts on ecological health of the river based on distribution of macroinvertebrate communities. As for the factors influencing the river health, multivariate statistical analysis exposed the influence of the metal concentration (total Fe), dissolved oxygen, and to a lesser degree, the nutrient enrichment.*

## TÓM TẮT

*Chương trình quan trắc là một yếu tố quan trọng trong quản lí chất lượng lưu vực sông. Tuy nhiên phân tích các số liệu quan trắc giúp xây dựng các giải pháp phục hồi và cải thiện chất lượng dòng sông là mục tiêu cuối cùng của quan trắc chất lượng. Nghiên cứu được thực hiện với 2 mục tiêu: (1) xác định các yếu tố quan trọng gây ảnh hưởng đến chất lượng nước lưu vực sông; (2) phân loại các điểm quan trắc dựa trên chất lượng hệ sinh thái. Các phép phân tích thống kê đa biến được ứng dụng nhằm giải quyết hai bài toán trên bao gồm tương quan tham biến và không tham biến, phân tích tương quan chuẩn, phân tích nhóm theo bậc dựa trên khoảng cách Bray-Curtis. Các phép phân tích thông kê trên được áp dụng nhằm đánh giá ảnh hưởng các hoạt động của con người đến chất lượng hệ sinh thái dựa trên phân bố các quần xã động vật không xương sống cỡ lớn (ĐVKXSCL). Kết quả nghiên cứu cho thấy các yếu tố ảnh hưởng quan trọng nhất là hàm lượng kim loại (tổng Fe), hàm lượng oxy hòa tan, và các chỉ tiêu liên quan đến ô nhiễm hữu cơ. Kết quả trên góp phần xây dựng các giải pháp phục hồi và bảo vệ dòng sông.*

## I. INTRODUCTION

Running water is the most important freshwater resource in Vietnam, being used for a variety of life purposes. In recent years, however the rapid socio-economic development has significantly affected the environment of river basins. Intensive agriculture, industrialisation and rapid urbanisation are the most important threats to deterioration of surface water quality. The maintenance of high quality running water has become an increasingly crucial issue in recent years, as greater demand has been placed on water resources [1].

In running water management, the monitoring of river water plays crucial role. The ecological quality of rivers and streams depends on their physical, chemical and biological properties. Therefore, there is now widespread recognition that not only chemical analyses but biological techniques are required for an appropriate assessment of river quality [2].

Assessment of river water quality based on monitoring data is more important task in river management. Recently, different techniques of data analysis are more and more applied to support river management. Data analysis can provide a better elucidation of river status in order to detect the cause of this status as well as improve assessment methods and consequently help to setup cost effective monitoring programs [3]. Applications of data analysis for information and decision support in river management are illustrated in Fig. 1.

Increasing computer power and the professional need to extract objective information from observed data have lead to complex databases. In tandem, statistical science had become a broad discipline with well developed methods and techniques for the design and analysis of wide range empirical studies.

The aim of the present study was to apply different multivariate statistical analysis to study monitoring data from the Du river basin in order to assess impacts of environmental stresses on aquatic life and consequently on river water quality.

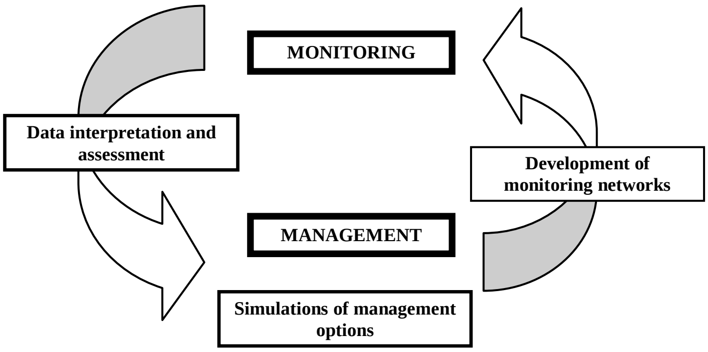


*Fig.1 Applications of data analysis in river management [4]*

## II. MATERIAL AND METHODS

*Study area*

The Du river is one of the main tributaries of the upstream part of the economically important Cau river, which is one of the most polluted river basins in Vietnam. The Du river is a medium-sized river with a length of 44 km and a catchment area of 360 km2. The Du runs across populated areas and receives impacts from domestic, agricultural and especially from industrial (metal mining and metal ore extraction) and coal mining activities at the Phan Me coal mine [5].

The 15 sites were sampled during seven campaigns from 2006 to 2008 and it was ascertained that wet and dry seasons were represented. The 19 physical-chemical and structural measurements were collected (Table 1).

Macroinvertebrates were collected by kick sampling and identified up to family level except Oligochaeta and Hydrachnida at class level by means of specific keys [6-9]. The Vietnamese version of the British Monitoring Working party (BMWP-Viet) and Average Score per Taxon (ASPT-Viet) was calculated based on BMWP-Viet scoring system [9]. The value of the index is the total of the scores of all taxa present in the list, where each family in the sample is counted once, regardless the number of species and individuals.

$$BMWP = \sum_{i=1}^{n} ToleranceScore_i$$

The BMWP index is divided by the number of scoring families present in the taxa list, the result is known as the Average Score Per Taxon (ASPT) index. The BMWP-Viet values were divided into classes each representing the ecological quality of the river water. Total taxa richness of Ephemeroptera, Plecoptera and Trichoptera (EPT) is also an important biotic index used in assessment of river health.

*Statistical analysis*

In studies exploring aquatic ecosystem, the correlation analysis is a powerful statistical tool for assessing the relationship between two variables [10]. Non-parametric correlation analysis based on Pearson product moment (*r*) was performed with each pair of variables collected from the monitoring campaigns. Any correlation with an absolute value of 0.2 or higher can be considered as a strong correlation [11].

Analysis of relationship between environmental variables and ecological scores was done by mean of Spearman ranking ($\rho$). The major advantages of using the non-parametric correlation methods over parametric are that the $\rho$ value is more flexible than Pearson product moment correlation ($r$) since it is based on ranks, is insensitive to outliers, requires no assumptions about normality, and can be used to assess the strength of monotonic (linear and nonlinear) associations [12]. The Spearman's rank was also applied to study correlations between biotic index BMWP-Viet and chemical variables.

To get an insight in the environmental parameter affecting macroinvertebrate community, the Canonical Correspondence Analysis (CCA) option from the program package PC-ord 4 was applied [13]. Prior to CCA analysis, all data were log transformed, except pH, which is already on a log scale. In addition, hierarchical cluster analysis using Bray-Curtis distance was also applied to assess the main reasons for clustering. The clusters were checked to which conditions of the habitat and anthropogenic impacts classes belonged.

## III. RESULTS AND DISCUSSION

*Environmental variables*

A basic set of environmental variables were registered in conjunction with the biological data (Table 1).

A correlation matrix (Table 2) between the hydromorphological and the physical-chemical characteristics reveals relatively strong correlations among hydromorphological variables (P/R class, stream width, depth and current). The strongest correlations were observed between the flow velocity and the stream depth ($r$= 0.61; $p$<0.001), between the altitude and the P/R class, stream width ($r$=-0.70 and $r$=-0.63 respectively; $p$<0.001) and between the pH and the total Fe concentration ($r$= -0.72; $p$<0.001).

*Table 1. River characteristics in Du river basin (2006-2008).*

| Variable | Unit | Min | Max | Mean | Std. Dev. |
|---|---|---|---|---|---|
| Elevation | m, a.s.l. | 106 | 380 | 217 | 81 |
| Pool/Riffle | category | 1 | 5 | 2.9 | 1.3 |
| Flow | m/s | 0 | 0.95 | 0.27 | 0.21 |
| Width | m | 0.63 | 28.9 | 7.51 | 6.54 |
| Depth | m | 0.05 | 0.95 | 0.29 | 0.16 |
| Dissolved oxygen | mg/l | 1.0 | 9.2 | 6.7 | 1.4 |
| pH | - | 3.10 | 8.36 | 6.93 | 1.15 |
| Conductivity | mS/m | 0 | 66 | 25 | 12 |
| Turbidity | NTU | 0 | 441 | 40 | 93 |
| Water temperature | $^{o}C$ | 14 | 36 | 25 | 5 |
| $BOD_5$ | mg/l | 2 | 20 | 8 | 3.7 |
| COD | mg/l | 2.7 | 82.1 | 23.7 | 15.2 |
| Kjeldahl N | mg N/l | 0.1 | 14.8 | 3.2 | 2.2 |
| $NO_2^-$-N | mg N/l | 0.01 | 0.78 | 0.08 | 0.13 |
| Total P | mg/l | 0.002 | 5.65 | 0.64 | 0.85 |
| $PO_4^{-3}$-P | mg P/l | 0.001 | 0.37 | 0.08 | 0.09 |
| $Cl^-$ | mg/l | 3.9 | 42.6 | 17.1 | 6.9 |
| Total Fe | mg/l | 0.014 | 27.00 | 2.30 | 4.04 |
| Total N sediment | mg/kg | 0.11 | 4.98 | 1.20 | 0.94 |

Strong negative correlation between the DO and the total Fe concentration ($r$= -0.64; $p$<0.001) revealed that iron ions in the water column may be an important consumption source of dissolved oxygen. Significant correlations between $BOD_5$ and pH ($r$= 0.4; $p$<0.001), and total Fe ($r$= -0.41; $p$<0.001) were explained by low $BOD_5$ level ($BOD_5$=2) registered in the Cat stream together with extremely low pH (down to 3.1) and high concentration of total Fe (up to 27 mg/l) in all monitoring campaigns. Acidification in the surface water may cause a transformation of organic matters represented by $BOD_5$ into inorganic substances.

Correlations between biological indices and chemical variables were computed using the non-parametric Spearman's rank coefficient of correlation ($\rho$).

The BMWP-Viet and EPT showed strong correlations with environmental variables in the Du river, while with the ASPT-Viet, significant correlations were only found with DO and the concentration of Fe.

*Table 2. Pearson correlation coefficients (r) calculated between the environmental variables. The presented correlation values are significant at p<0.05 (N=104).*

| | Altitude | Flow | Width | Depth | DO | pH | Conduc | WTemp | $BOD_5$ | COD | Kjd N | Total P |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Flow | -0.48 | -- | | | | | | | | | | |
| Width | -0.63 | 0.52 | -- | | | | | | | | | |
| Depth | -0.48 | 0.6 | 0.58 | -- | | | | | | | | |
| P/RClass | -0.7 | 0.33 | 0.51 | 0.47 | | | | | | | | |
| pH | | 0.32 | | 0.41 | 0.37 | -- | | | | | | |
| WTemp | | | | | | -0.25 | | -- | | | | |
| $BOD_5$ | | | | | | 0.4 | | | -- | | | |
| COD | | 0.35 | | 0.25 | 0.29 | 0.27 | | | | -- | | |
| Kjd N | | | | | | | | 0.53 | | | -- | |
| $NO_2$ | | | | 0.26 | | | | -0.34 | | 0.23 | | |
| Total P | | | | | | | | | | | 0.31 | |
| $P\text{-}PO_4^{3-}$ | | | | | 0.26 | | -0.3 | | | | | -- |
| Cl | -0.28 | 0.32 | | 0.27 | | | | | | | | -0.24 |
| Total Fe | | | | | -0.64 | -0.72 | | | -0.41 | -0.23 | | |
| TN Sed | 0.46 | | | | | | | | | | | |

*Table 3. Spearman rank correlations between biotic indices and environmental variables in the Du river (n=104).*

| | P/R Class | Altitude | Width | DO | BOD5 | Total Fe |
|---|---|---|---|---|---|---|
| BMWP Viet | -0.39** | 0.43** | -0.23* | 0.51** | 0.34** | -0.54** |
| ASPT Viet | n/c | n/c | n/c | 0.33* | n/c | -0.27** |
| EPT | -0.43** | 0.29** | n/c | 0.48** | 0.32** | -0.45** |

*$p$<0.05; ** $p$< 0.001; n/c: no significant correlation

*Site clustering*

CCA was applied for site clustering. Sites 10 and 11 are extremely polluted sites with a pH of 3.14 to 5.52. No taxa were found at site 11, while only Gerridae, Tabanidae and Planorbidae were found at site 10. These sites were classified very poor and excluded from further analysis to avoid impacts of outliers. Based on the results of the CCA, the sites were clustered into three groups: a group containing sites in good condition, a group containing sites in poor condition and an intermediate group containing sites with a moderate condition (Fig.1). The clustering was confirmed by clustering based on the BMWP-Viet values.

CCA performed on the whole dataset showed that the site clustering was based on physical variables including altitude, flow, depth, width and P/R class and chemical variables TNSed, DO, P-$PO_4$ and total Fe (significant at $r^2$=0.15). The Eigenvalues of the first and the second axis were 0.094 and 0.077, respectively (Fig. 2).

Hierarchical cluster analysis received agglomerative coefficient of 0.85. Dendogram of the cluster analysis based on Bray-Curtis (Fig.3) showed that samples could be clustered into several groups. In the first division, the samples from extremely polluted site 10 in the Cat stream were separated from remaining samples. In the second division, samples from sites at down stream with high impacts were split off. In the next division, samples were split off based on location altitude and then by impacts. The hierarchical cluster also confirmed clustering by CCA.

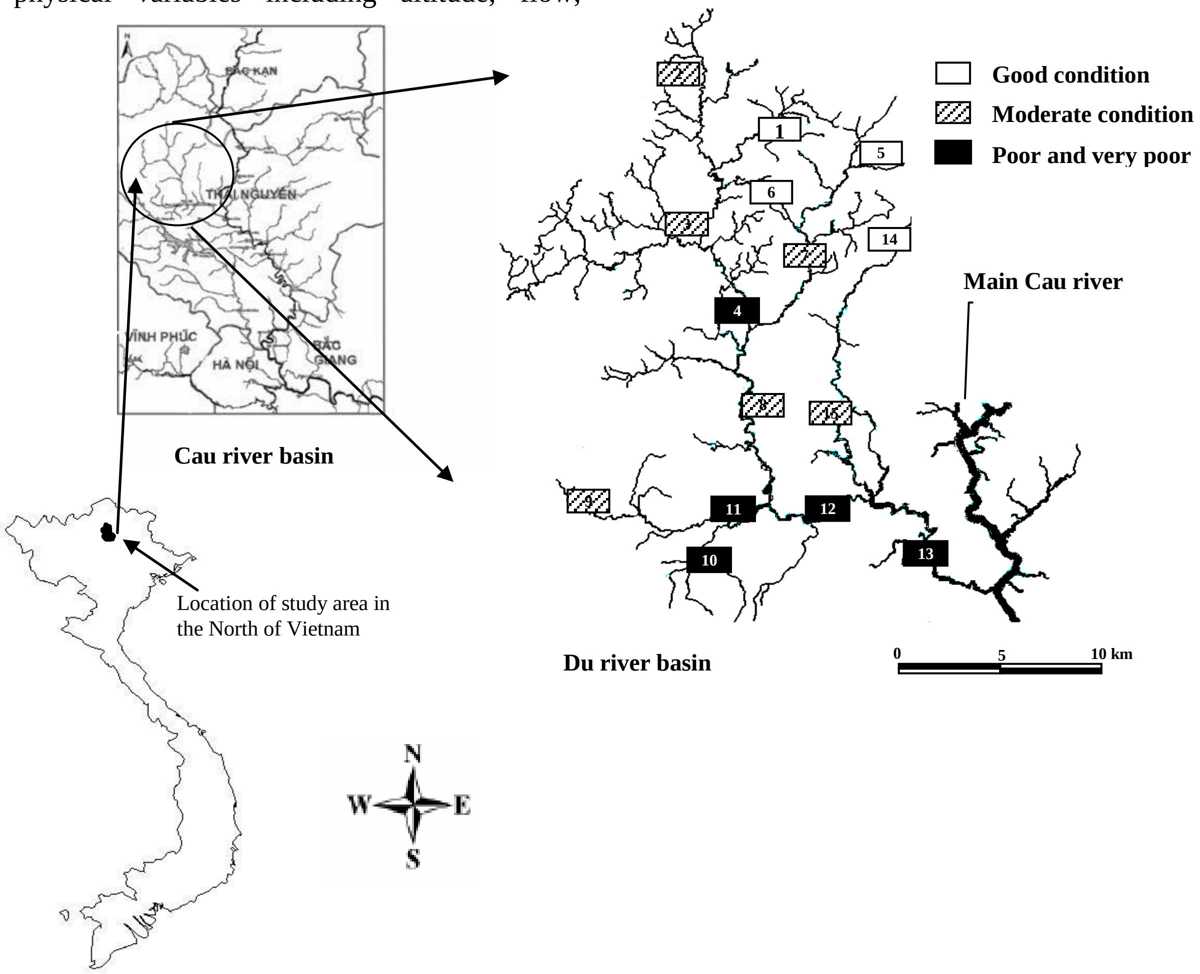


*Fig. 1 Location of the sampling stations in the Du river basin, with indication of the three identified clusters.*

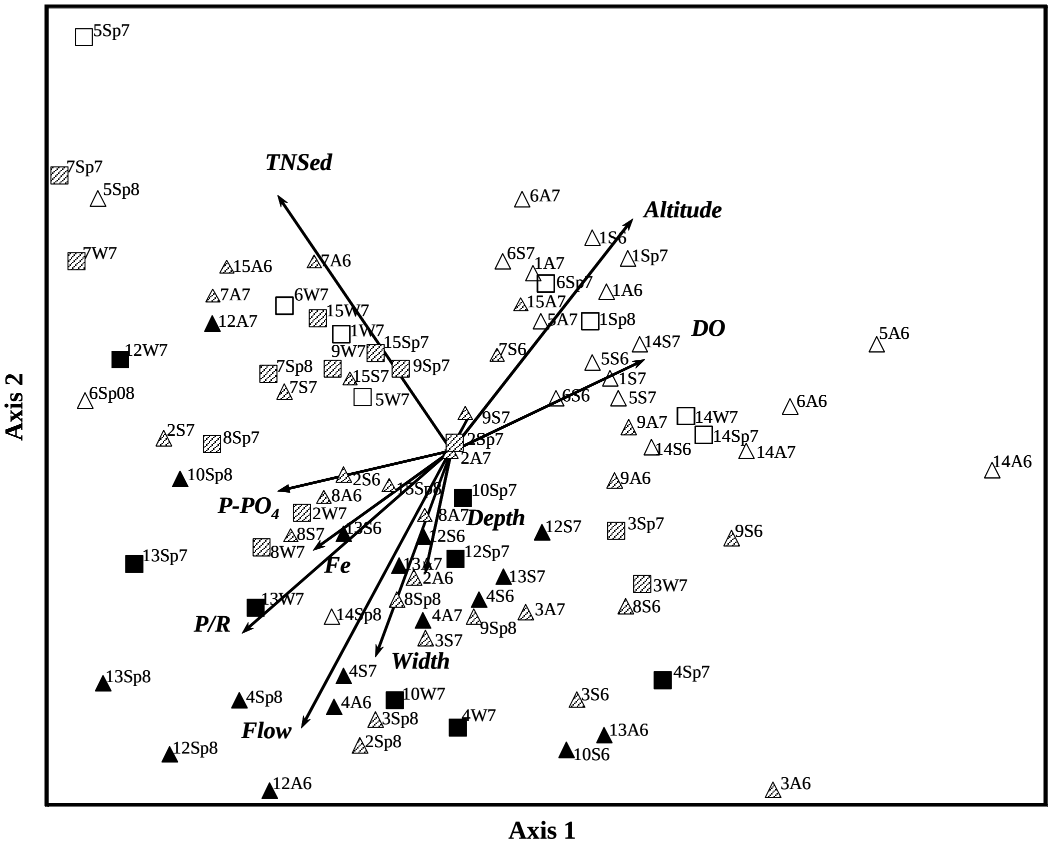


*Fig.2 Biplot of the sample scores and the environmental variables with indication of the three identified clusters (black: poor condition; dashed: moderate condition; white: good condition) and the sampling season (dry season: squares; wet season: triangles). Labels indicate the sampling site, the season (Sp: spring, S: summer, A: autumn, W: winter) and the year. Cut-off $r^2$ value: 0.15.*

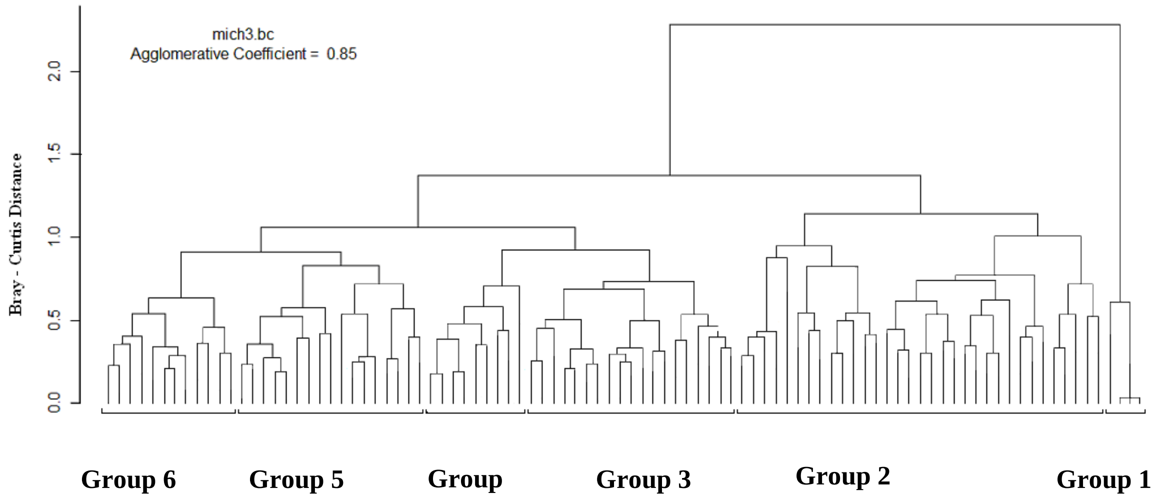


*Fig.3 Dendogram of hierarchical cluster analysis based on Bray-Curtis distance using square root transformed macroinvertebrate data from Du river basin Group 1: extremely polluted; Group 2: Down stream sites with high impacts; Group 3: Sites with low impact at low altitude; Group 4: Sites with moderate impacts at low altitude; Group 5: Sites with moderate impacts at high altitude; Group 6: Sites with low impacts at high altitude.*

Although the water quality characteristics of the studied sites did not exceed the norms for fresh watercourses (TCVN 5942-2005: Surface water quality standard), the diversity in the Du river basin was seriously affected by human activities. Wastewater from metal mining and metal extraction activities in some tributaries is the most important pollution source that caused a low pH and high metal concentrations in some sites. The basin is located in a highly populated area with intensive agriculture. The use of fertilizers led to high levels of nitrogen and phosphorous concentration in the river water.

Coal mining from the Phan Me open pits located downstream was also an important source of pollutants. The results of study showed clear impacts of anthropogenic activities in the Du river basin on distribution of macroinvertebrate and consequently ecological health of the river. Highly significant correlations were found between biotic indices and environmental variables such as $BOD_5$, DO and Total Fe (Table 3), which proved the sensitivity of these indices to environmental stress.

All statistical analysis applied in this study proved to be able to detect factors influencing distribution of macroinvertebrate communities and consequently ecological health of the river. Canonical Correspondence Analysis showed the most robustness in determining environmental factors responsible of these patterns.

The results proved interesting results for management purposes. Different multivariate statistical analysis methods confirmed that the driving variables for ecological health of the river were (1) pH, (2) $BOD_5$, (3) $N\text{-}NO_2$, (4) total Fe, (5) DO. Driving variables revealed important pollution sources of the Du river.

- Acid mine wastewater caused low pH and high metals concentration in water column.
- Untreated domestic wastewater discharged into the river caused excessive nutrient and high BOD level
- Runoff from agricultural land caused excessive nutrient load.
- River bed was destroyed by uncontrolled sand extraction in the downstream sites.

Based on these finding, immediate measures are suggested to restore the degraded parts and to prevent the deterioration of the pristine upstream parts such as (1) Acid mine wastewater treatment; (2) domestic wastewater collection and treatment at small and medium sizes; (3) Riparian buffer zone; (4) Improve agriculture management practice.

However, restoration actions should integrate technical and management issues, especially awareness raising for involvement of general public to the conservation of the river ecosystem.

## IV. CONCLUSIONS

The present study investigates the relationships between environmental variables and ecological health of the rivers based distribution of macroinvertebrate communities. Data analysis revealed that multivariate statistical analysis was a powerful tool in studying these relationships and that makes the method an efective tool for river management.

Monitoring strategies should focus on key variables identified by data analysis, while keeping physical variables as a key factor determining favourable habitat conditions for macroinvertebrates. Comparison between monitoring results in sites with the same physical habitat conditions can present meaningful interpretation on the pollution status of the sites. Further exploration on decision support systems is necessary to link monitoring results with the selection of appropriate restoration measures for the improvement of ecological river quality.

## REFERENCES

1. VEPA (2006). Environmental report of Vietnam 2006: The state of water environment in three river basins of Cau, Nhue-Day and Dong Nai river system; VEPA, Hanoi, Vietnam, 93pp.

2. *Wright J. F.* (1995); Development and use of a system for predicting the macroinvertebrate fauna in flowing waters; Freshwater Biology, 20: 181-197.

3. *Vanrolleghem P. A., Schilling W., Rauch W., Krebs P. and Aalderink H.* (1999); Setting up measuring campaigns for integrated wastewater modelling; Water Science & Technology, 39(4): 257-268.

4. *Goethals P. and De Pauw N.* (2001); Development of a concept for integrated ecological river assessment in Flanders, Belgium. Journal of Limnology 60(1): 7-16.

5. Thai Nguyen DONRE (2006); Environmental status in Thai Nguyen provinces 2004- 2005; Thai Nguyen DONRE, Thai Nguyen, Vietnam 138pp.

6. *Bouchard R. W. J.* (2004); Guide to aquatic macroinvertebrates of the Upper Midwest; Water Resources Center, University of Minnesota, St. Paul, USA, 208pp.

7. *Dudgeon D.* (1999); Tropical Asian streams. Zoobenthos, ecology and conservation; Hongkong University press, Hongkong, 830pp.

8. *McCafferty W. P. and Provonsha A. V.* (1983); Aquatic Entomology: The fishermen's and ecologists' illustrated guide to insects and their relatives; Jones and Bartlett Publishers, Boston, USA, 448pp.

9. *Nguyen X. Q., Mai D. Y., Pinder C. and Tilling S.* (2004); Biological surveillance of freshwaters: A practical manual and identification key for use in Vietnam; Vietnam National University Publishers, Hanoi, Vietnam, 109pp

10. *Gotelli N. J. and Ellison A. M.* (2004); Aprimr of ecological statistics; Sinauer Associates Inc. Sunderland, Massachusetts. 510pp

11. *Legendre P. and Legendre L.* (1998); Numerical Ecology. 2nd English Edition; Elsevier Science, Amsterdam, The Netherlands, 870pp.

12. *Sutherland R. A.* (2001); Analysis and commentary on “Statistical methods and pitfalls in environmental data analysis” by Yue Rong; Environmental Forensics, 2: 265-274.

13. *McCune B. and Mefford M. J.* (1999); Multivariate analysis for ecological data Version 4.0; MjM Software, Oregon.

---

*Contact:* *Hoang Thu Huong* - Tel: (+844)3868.1686, Email: huonght-inest@mail.hut.edu.vn
Hanoi University of Technology - Institute for Environmental Science and Technology
No. 1, Dai Co Viet road, Hanoi, Vietnam
*Peter Goethals* - Tel.: (+32) 09/264.37.68 , Email: Peter.Goethals@UGent.be
Ghent University - Faculty of Bioscience Engineering
J. Plateaustraat 22 - 9000 Gent