MULTIVARIATE STATISTICAL TECHNIQUES FOR THE ASSESSMENT OF SURFACE WATER QUALITY OF FUJI RIVER BASIN, JAPAN

Sangam SHRESTHA[1] and Futaba KAZAMA[2]

[1]Member of JSCE, PhD Student, Department of Ecosocial System Engineering, University of Yamanashi
(4-3-11, Takeda, Kofu, Yamanashi, 400-8511, Japan)
E-mail:sangam@ccn.yamanashi.ac.jp
[2]Member of JSCE, Associate Professor, Department of Ecosocial System Engineering, University of Yamanashi
(4-3-11, Takeda, Kofu, Yamanashi, 400-8511, Japan)
E-mail:kfutaba@yamanashi.ac.jp

Different multivariate statistical techniques were used to evaluate temporal and spatial variations of surface water-quality of Fuji river basin using data sets of 8 years monitoring at 13 different sites. The hierarchical cluster analysis grouped thirteen sampling sites into three clusters i.e. relatively less polluted (LP), medium polluted (MP) and highly polluted (HP) sites based on the similarity of water quality characteristics. The principal component analysis/factor analysis indicated that the parameters responsible for water quality variations are mainly related to temperature (natural), organic pollution (point sources) in LP areas; organic pollution (point sources) and nutrients (non point sources) in MP areas; and organic pollution and nutrients (point sources) in HP areas. The discriminant analysis showed that five water quality parameters account for most of the expected temporal variations whereas six water quality parameters account for most of the expected spatial varitations in surface water quality of Fuji river basin.

Key Words : *Fuji river,water quality,cluster analysis, factor analysis, principal component analysis, discriminant analysis*

1. INTRODUCTION

A river is a system comprising both the main course and the tributaries, carrying on one-way flow of significant load of constituents in dissolved and particulate phases from both natural and anthropogenic sources. The quality of river at any points reflects several major influences, including the lithology of the basin, atmospheric inputs, climatic conditions and anthropogenic inputs. On the other hand, rivers play a major role in assimilation or carrying off the municipal and industrial wastewater and run-off from agricultural land. Seasonal variations in precipitation, surface runoff, groundwater flow and pumped in and outflows have a strong effect on river discharge and subsequently on the concentration of pollutants in river water[1)]. Therefore the effective and long term management of rivers requires a fundamental understanding of major factors affecting the change in water quality parameters. However, due to spatial and temporal variations in surface water quality, a monitoring program that produces a large and complex data set is required to provide a representative and reliable estimation of the water quality. Therefore, it is often difficult to interpret and drawing meaningful conclusions from such a large and complex dataset[2)]. The use of different multivariate statistical techniques such as cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA) helps for better interpretation of these complex data matrices to understand the water quality of the studied systems. It allows to identify the possible factors/sources that influence the water systems and offers a valuable tool for management of water resources as well as rapid solutions on pollution problems[3),4)]. The multivariate statistical techniques has been applied to characterize and evaluate surface water quality and also proved to be useful for evidencing temporal and spatial variations caused by natural and anthropogenic factors related to seasonality [5),6)] .

In this study, a large data matrix of 8-year (1995-2002) monitoring program are subjected to different multivariate statistical techniques in order a) to extract information about the similarities or dissimilarities between sampling sites, b) to identifiy water quality variables responsible for spatial and temporal variations in river water quality, c) to extract the hidden factors explaining the structure of the database, and d) to know the influence of possible sources (natural and anthropogenic) on the surface water quality of Fuji river basin.

2. METHODOLOGY

(1) Study area

The study area, Fuji river basin, drained by the Fuji river is located in the central part of Japan (**Fig.1**). The basin area is 3,570 km^2 and the mainstream length is 128 km.The river originates as the Kamanashi river from Mt. Komagatake in the north of the Southern Alps, and as the Fuefuki river from the north of Yamanashi prefecture. These two river joins and flows together in the south of Kofu basin as the Fuji river and subsequently flows to the Pacific Ocean. The average flow of Kamanashi river at Funayamabashi is about 10 m^3/s; Fuefuki river at Torinkyo is about 20 m^3/s and of Fuji river at Fujibashi is about 72 m^3/s. These rivers drain the major rural, agricultural, urban and industrial areas of Yamanashi prefecture and discharge into Suruga Bay, Pacific Ocean. The Fuji river is the major source for agriculture and industrial activities located in the downstream areas. The geological features of the basin are very complex and fragile. The basin has more than 75% area covered by forest landuse type. The forest land is distributed mostly in mountainous area whereas agriculture and grassland areas are sparsely distributed throughout the basin. The orchard plantations and urban areas are mostly situated along the water bodies. The summers are hot (av. 24^{0}C) and humid and winters are cold (av. 3^{0}C). The basin receives a mean annual precipitation of approximately 2,100 mm.

(2) Monitored parameters

The data sets of thirteen water quality monitoring stations (listed in **Fig.1**) comprising eleven water quality parameters monitored monthly during 8 years (8 x 13 x 11 x 12 = 13,728 observations) were obtained from the Environment Division of Yamanashi Prefecture (EDYP), Japan[7].The selected water quality parameters include water temperature (^{o}C), dissolved oxygen (mgl^{-1}), 5-days biochemical oxygen demand (mgl^{-1}), chemical oxygen demand (Mn) (mgl^{-1}), pH, total suspended solids (mgl^{-1}), electrical conductivity (µScm^{-1}), total coliforms (MPN/100ml), nitrate nitrogen (mgl^{-1}), ammonical nitrogen (mgl^{-1}), and inorganic dissolved phosphorus (mgl^{-1}).

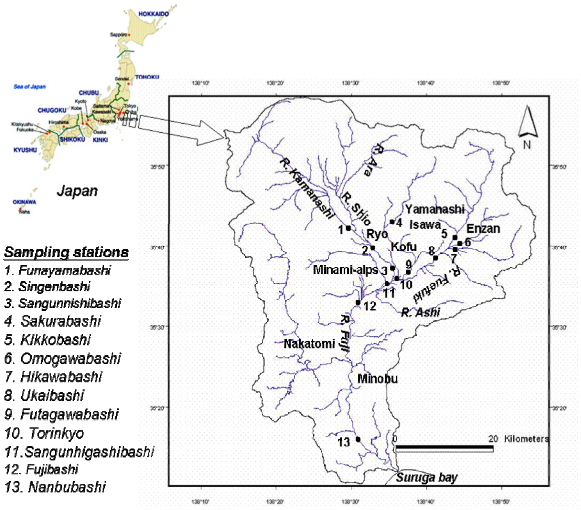

Fig. 1 Water quality monitoring stations in the Fuji river basin

(3) Data treatment and multivariate statistical methods

The river water quality data sets were subjected to four multivariate techniques: cluster analysis (CA), principal component analysis (PCA), factor analysis (FA) and discriminant analysis (DA). DA was applied on raw data[6] whereas PCA, FA and CA were applied on experimental data standardized through *z*-scale transformation in order to avoid misclassifications arising from the different orders of magnitude of both numerical value and variance of the parameters analyzed[8],[4]. The average values of eight years were taken for each water variables in cluster analysis. All the mathematical and statistical computations were made using Microsoft Office Excel 2003 and STATISTICA 6.

a) Cluster analysis

Cluster analysis is an unsupervised pattern recognition technique that uncovers intrinsic structure or underlying behaviour of a data set without making apriori assumption about the data, in order to classify the objects of the system into categories or clusters based on their nearness or similarity[1]. In this study, hierarchical agglomerative CA was performed on the normalized data set using the Ward's method and Euclidean distances as a measure of similarity. The Ward's method uses an analysis of variance approach to evaluate the distances between clusters attempting to minimize the sum of squares of any two clusters that can be formed at each step. The spatial variability of water quality in the whole river basin was determined from CA, using the linkage distance reported as $D_{\text{link}}/D_{\max}$, which represent the quotient between the linkage distances for a particular case divided by the maximal linkage distance[3].

b) Principal compenent analysis /factor analysis

The PCA is designed to transform the original variables into new, uncorrelated variables (axes) called the principal components, that are linear combinations of the original variables. The new axes lie along the directions of maximum variance. The PC provides information on the most meaningful parameters, which describes whole data set affording data reduction with minimum loss of original information [5].

The FA follows PCA. The main purpose of FA is to reduce the contribution of less significant variables in order to simplify even more of the data structure coming from PCA. This purpose can be achieved by rotating the axis defined by PCA, according to well established rules, and constructing new variables, also called varifactors (VF). The PC is a linear combination of observable water quality variables whereas VF can include unobservable, hypothetical, latent variables[1),5]. The PCA of the normalized variables was performed to extract significant PCs and to further reduce the contribution of variables with minor significance; these PCs were subjected to varimax rotation (raw) generating VFs [9].

c) Discriminant analyis

In Discriminant Analysis, multiple quantitative attributes are used to discriminate between two or more naturally occurring groups. In contrast to CA, DA provides statistical classification of samples and it is performed with prior knowledge of membership of objects to a particular group or cluster. Further, DA helps in grouping the samples sharing common properties. This technique constructs a discriminant function for each group as in equation below:

$$f\left(G_i\right)=K_i+\sum_{j=1}^{n}W_{ij}\,p_{ij} \qquad (1)$$

where i is the number of groups (G), K_i is the constant inherent to each group, n is the number of parameters used to classify a set of data into a given group, W_j is the weight coefficient, assigned by DA to a given selected parameter (P_j) and P_j the analytical value of the slected parameter. The weight coefficients maximizes the distance between the means of the criterion (dependent) variable. The classification table, also called a confusion, assignment, or prediction matrix or table, is used to assess the performance of DA. This is simply a table in which the rows are the observed categories of the dependent and the columns are the predicted categories of the dependents.When prediction is perfect, all cases will lie on the diagonal. The percentage of cases on the diagonal is the percentage of correct classifications. In this study, four groups for temporal (four seasons) and three groups for spatial (three sampling regions) evaluations have been selected and the number of analytical parameters used to assign a measure from a monitoring site into a group (season or monitoring area).DA was performed on each raw data matrix using standard, forward stepwise and backward stepwise modes to construct discriminant functions to evaluate both the spatial and temporal variations in river water quality of the basin. The site (spatial) and the season (temporal) were the grouping (dependent) variables, whereas all the measured parameters constituted the independent variables.

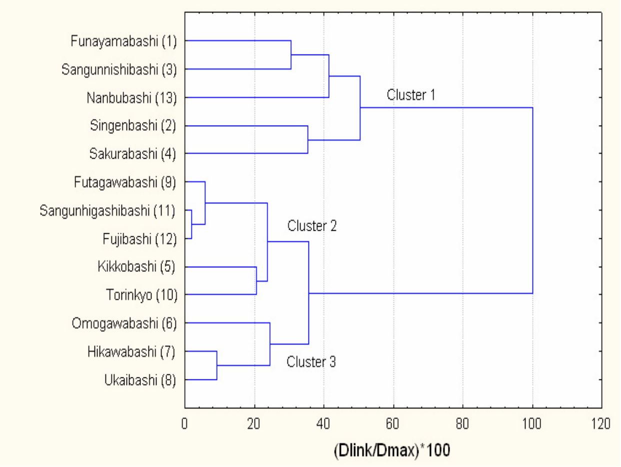

Fig. 2 Dendrogram showing clustering of sampling sites in the Fuji river basin

3. RESULTS AND DISCUSSIONS

(1) Spatial similarity and site grouping

The cluster analysis was used to detect the similarity groups between the sampling sites. It rendered a dendrogram (**Fig.2**), grouping all the thirteen sampling sites of the basin into three statistically significant clusters at ($D_{\text{link}}/D_{\max}$) x100<60. These three groups have the similar characteristic features and natural background source types. Since we used the hierarchical agglomerative cluster analysis, the number of clusters was also decided by practicality of the results as there is ample information (e.g. landuse, location of wastewater treatment plants etc) available about the study sites.

The cluster 1 (Sakurabashi, Singenbashi, Nanbubashi, Sangunnishibashi and Funayamabashi) correspond to relatively less polluted (LP) sites. In cluster 1, four stations, Sakurabashi, Singenbashi, Sangunnishibashi and Funayamabashi are situated at the upstream sites and Nanbubashi is situated at the most downstream site of the river. The inclusion of the most downstream sampling location, Nanbubashi, in cluster 1 group suggests the self purification and assimilative capacity of the river.The cluster 2 (Torinkyo, Kikkobashi, Fujibashi, Sangunhigashibashi, and Futagawabashi) correspond to highly polluted sites (HP). These stations receive the pollution mostly from domestic wastewaters, waste water treatment plants and industrial effluents located in the city areas (Kofu, Yamanashi, Isawa). The cluster 3 (Ukaibashi, Hikwabashi and Omogawabashi) correspond to moderately polluted (MP) sites and these stations receive the pollution from nonpoint sources i.e. mostly from agricultural and orchard activities. While the analysis was done from the samples taken only during low flow period, it can be said that there is a groundwater contribution in pollution loading into the water bodies of this area. These results indicate that CA technique is useful in offering reliable classification of surface water in the whole region and will make it possible to design a future spatial sampling strategy in an optimal manner which can reduce the number of sampling stations and cost associated with it.

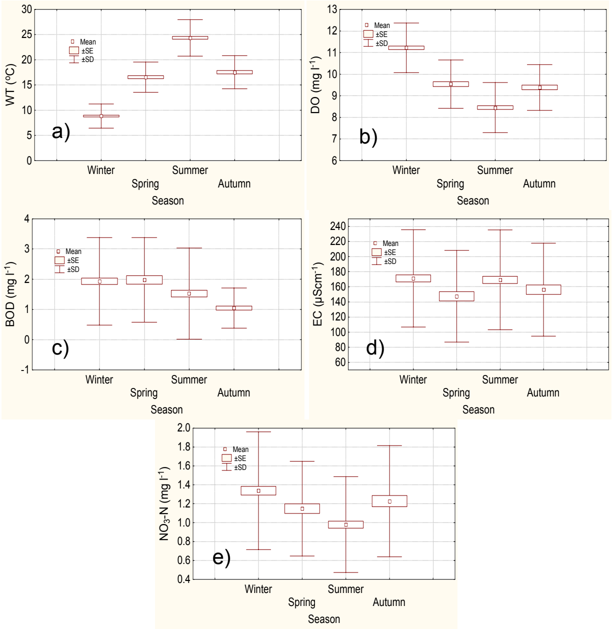

Fig. 3 Temporal variations : a) temperature ; b) DO ; c) BOD; d) EC; e) NO$_3$-N

(2) Temporal and spatial variations in river water quality

The temporal variations of the river water quality parameters were evaluated through DA. Temporal DA was performed on raw data after dividing the whole data set into four seasonal groups (spring, summer, autumn and winter). In forward stepwise mode, variables are included step-by-step beginning with the more significant until no significant changes are obtained, whereas, in backward stepwise mode, variables are removed step-by-step beginning with the less significant until no significant changes are obtained. Both the standard and forward stepwise mode, discriminant functions (DF) using 11 discriminant variables, respectively, rendered the corresponding classification matrix (CM) assigning 85% cases correctly . However, in backward stepwise mode DA gave CMs with 85% correct assignations using only five discriminant parameters. Thus, the temporal DA results suggest that temperature, dissolved oxygen, biochemical oxygen demand, electrical conductivity and nitrate nitrogen are the most significant parameters to discriminate between the four different seasons, which means that these five parameters account for most of the expected temporal variations in the river water quality.

As identified by DA, box and whisker plots of the selected parameters showing seasonal trends are given in **Fig. 3**. The water temperature is lowest during winter season and highest during summer season. The dissolved oxyen is observed to be increased from spring season to winter season. The mean concentration of electrical conductivity is higher in summer and winter season as compared to spring and autumn seasons. However, decrease in nitrate nitrogen concentration from winter to summer followed by increase in autumn season is observed. Similar temporal variations in nitrate nitrogen are also reported in this area[10]. Further, decrease in the concentration of biochemical oxygen demand from winter to autumn is also observed in the basin.

Spatial DA was performed with the same raw data set comprised of 11 parameters after grouping into three major classes of LP, MP and HP sites as obtained through CA. The site (clustered) was the grouping (dependent) variable, while all the measured parameters constituted the independent variables. Both the standard and forward stepwise mode DFs using 11 discriminant parameters rendered the corresponding CMs assigning more than 83% cases correctly . Whereas the backward stepwise mode DA gave CMs with the same percentage (81%) correct assignations using only six discriminant parameters. Backward stepwise DA shows that temperature, biochemical oxygen demand, pH, electrical conductivity, nitrate nitrogen and ammonical nitrogen are the most significant discriminating parameters in space.

Box and whisker plots of discriminating parameters identified by spatial DA (backward stepwise mode) were constructed to evaluate different patterns associated with spatial variations in

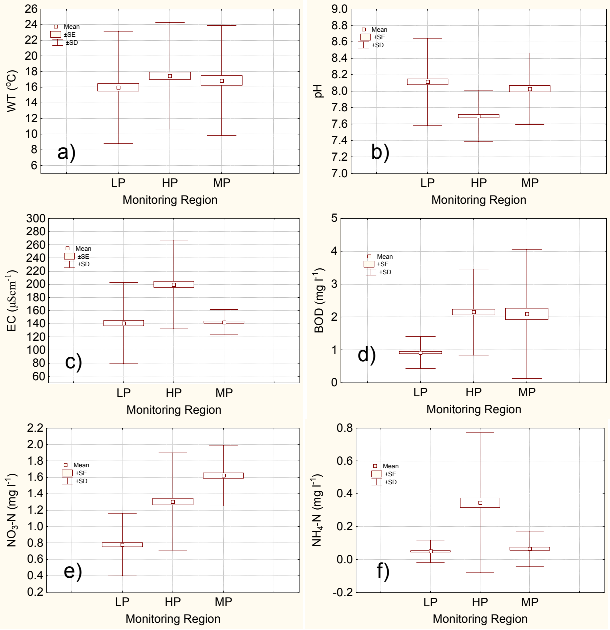

Fig. 4 Spatial variations: a) temperature; b) pH; c) EC; d) BOD; e) NO$_3$-N and f) NH$_4$-N

river water quality (**Fig. 4**). The river water temperature is highest in HP sites as they receive the discharge from domestic wastewaters, waste water treatment plants, and industrial effluents located in city areas that increase the water temperature.The trends for BOD, pH, EC and NH$_4$-N suggests for high load of dissolved organic matter in the HP sites added from the domestic wastewaters, wastewater treatment plants and industrial effluents located at the upstream areas of the monitoring stations.This results in the anaerobic condition in the river which results in formation of ammonia and organic acids.The hydrolysis of these acidic materials causes a decrease of pH in these sites. The highest mean concentration of nitrate nitrogen is observed in MP sites. This can be attributed to the use of nitrogenous fertilizer in the orchard and agricultural areas. The study conducted in this area[10] also supports that the orchard and agriculture related activities are the source of nitrate nitrogen in these areas.

(3) Data structure determination and source identification

Principal component analysis/factor analysis was performed to the normalized data sets (11 variables) separately for the three different regions viz. LP, MP and HP, as delineated by CA techniques, to compare the compositional pattern between analyzed water samples and to identify the factors that influence each one. The input data matrices (variables x cases) for PCA/FA were [11 x 218] for LP and HP and [11 x 127] for MP sites. PCA of the three data sets evolved five PCs for LP and MP sites and three PCs for HP sites with eigen value >1, explaining 75.24, 77.61 and 65.39% of the total variance in respective water quality data sets. An eigenvalue gives a measure of the significance of the factor: the factors with the highest eigenvalues are the most significant. Eigenvalues of 1.0 or greater are considered significant[11]. Equal numbers of VFs were obtained for three sites through FA performed on the PCs. Only the strong factor loading (>0.75)[8] was considered to explain the data structure in this study.

For the data set pertaining to LP sites, among five VFs, the VF1 explaining 22.75% of total variance has strong positive loading temperature and strong negative loadings on DO. The VF1 represent the seasonal effects .VF2 explaining 18.59% of the total variance has negative loadings of suspended solids and chemical oxygen demands. This factor explains the erosion from upland areas during rainfall events and the positive correlation with COD indicates the loading of partially decayed organic matters from forested areas. VF3 and VF4 explaining 14.09% and 10.47 respectively of total variance has strong positive loadings on ammonical nitrogen and nitrate nitrogen. This factor represents the organic and inorganic pollution from domestic wastes. VF5 explaining the lowest variance (9.31%) has strong positive loadings on pH and represents the physiochemical source of variability.

For the data set representing the MP sites, among total five significant VFs, the VF1 explaining about 24.98% of total variance has strong positive loadings on nitrate nitrogen and biochemical oxygen demand. This factor represents the contribution of nonpoint source pollution from orchard and agricultural areas. This fact is also supported by the study conducted in Yamanashi Prefecture[10),12)]. VF2, explaining about 20.20% of total variance, has strong positive loading on temperature and strong negative loadings on dissolve oxygen. This factor can be attributed to the seasonal change. VF3, explaining about 13.51% of total variance has strong positive loadings on pH and phosphate phosphorus. VF4, explaining 9.60% of total variance has strong positive loadings on total suspended solids and chemical oxygen demands. This factor represents the erosion effect during the cultivation of soil and associated organic matter. VF5 has strong negative loadings on electrical conductivity. This factor suggests the dilution effect.

Lastly, for the data set pertaining to water quality in HP sites, among the three VFs, VF1 explaining 32.83% of total variance has strong positive loadings on biochemical oxygen demand, chemical oxygen demand, electrical conductivity, ammonical nitrogen and phosphate phosphorus. This organic factor can be interpreted as representing influences from point source such as of discharges from wastewater treatment plants, domestic wastewaters and industrial effluents. VF2, explaining 17.59% of total variance, has negative loadings on pH. The strong negative loading in pH is due to the anaerobic condition in the river due to the loading of high dissolve organic matter which results in formation of ammonia and organic acids leading to decrease in pH. VF3, explaining 14.97% of total variance has strong positive loadings on temperature. This factor represents the seasonal effect of temperature.

4. CONCLUSION

In this case study, different multivariate statistical techniques were used to evaluate the temporal and spatial variations in surface water quality of Fuji river basin. The hierarchical cluster analysis grouped thirteen sampling sites into three clusters of similar water quality characteristics. Based on the information obtained, it is possible to make the design of a future sampling strategy in an optimal way that could reduce the number of sampling stations and the cost incurred in it. Although the factor analysis/principle component analysis did not result much in considerable data reduction, but it helped to extract and identify the factors/sources responsible for river water quality variations in three different sampling sites. The discriminant analysis rendered an important data reduction when compared with factor analysis/ principal component analysis as it uses only five parameters (temperature, dissolved oxygen, biochemical oxygen demand, electrical conductivity and nitrate nitrogen) affording more than 85% correct assignations in temporal analysis, while six parameters (temperature, biochemical oxygen demand, pH, electrical conductivity, nitrate nitrogen and ammonical nitrogen) affording more than 81% correct assignations in spatial analysis of three different sampling sites of the basin. In other words, only 5 parameters are the most important water quality parameters that differs significantly among four seasons. Similarly, only 6 water quality parameters are the most important water quality parameters that differ significantly among three monitoring regions. From these results we can say that sampling of few parameters can also reflect the temporal and spatial variations that can be considered for the future water quality studies in the study basin.Thus, this study illustrates the usefulness of multivariate statistical techniques to analyse and derive interpretation from complex data set for water quality assessment. Furthermore, it aids in identification of pollution sources/factors, and understanding of temporal and spatial variations of water quality for the effective river water quality management.

ACKNOWLEDGEMENTS: The authors would like to extend sincere thanks to Ms. Yuki Hiraga for her help in the database development and The Fuji Xerox Setsutaro Kobayashi Memorial Fund for providing funding support to conduct this study.

REFERENCES

1) Vega, M., Pardo, R., Barrado, E., Deban, L.: Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Res*,Vol. 32, pp. 3581–3592, 1998.

2) Dixon, W., Chiswell, B.: Review of aquatic monitoring program design. *Water Res.* Vol. 30, pp. 1935–1948, 1996.

3) Wunderlin, D.A., Diaz, M.P., Ame, M.V., Pesce, S.F., Hued,A.C., Bistoni, M.A.:. Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquia river basin (Cordoba-Argentina). *Water Res,* Vol. 35, pp. 2881–2894, 2001.

4) Simeonov, V., Stratis, J.A., Samara, C., Zachariadis, G.,Voutsa, D., Anthemidis, A., Sofoniou, M., Kouimtzis,T.: Assessment of the surface water quality in Northern Greece. *Water Res,*Vol. 37, pp. 4119–4124, 2003.

5) Helena, B., Pardo, R., Vega, M., Barrado, E., Ferna´ ndez, J.M.,Ferna´ ndez, L.: Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga river, Spain) by principal component analysis. *Water Res,* Vol. 34, pp. 807–816, 2000.

6) Singh, K.P., Malik, A., Sinha, S.: Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques- a case study. *Anal. Chim. Acta*, Vol. 538, pp. 355–374, 2005.

7) EDYP: Result of Water Quality Measurement: Public and Ground water. Atmospheric Water Quality Control Section, Forest and Environment Division of Yamanashi Prefecture, 2004.

8) Liu, C.W., Lin, K.H., Kuo, Y.M.: Application of factor analysis in the assessment of groundwater quality in a Blackfoot disease area in Taiwan. *Sci. Tot. Environ*. Vol. 313, pp.77–89,2003.

9) Brumelis, G., Lapina L., Nikodemus, O., Tabors, G.: Use of an artificial model of monitoring data to aid interpretation of principal component analysis. *Environmental Modelling and Software,* Vol. 15, pp. 755-763, 2000.

10) Fukasawa, E.: Determination of origin of nitrate nitrogen in Fuefuki river using stable isotope method. Bachelor Thesis, Department of Ecosocial System Engineering, University of Yamanashi, Japan, 2005.

11) Kim, J.-O., Mueller, C.W.: Introduction to Factor Analysis: What It is and How to Do It Quantitative Applications in the Social Sciences Series. Sage University Press, Newbury Park, 1987.

12) Kazama, F., Yoneyama, M.: Nitrogen generation in the Yamanashi prefecture and its effects on the groundwater pollution: *Int. Environmental science*, Vol.15, pp. 293-298, 2002.

(Received September 30, 2005)