

Extraction of a Parameter as an Index to Assess Water Quality of the Tamagawa, Tokyo, Japan, by Using Neural Networks and Multivariate Analysis

Junko KAMBE^a, Tomoo AOYAMA^b, Aiko YAMAUCHI^c and Umpei NAGASHIMA^{d*}

^aFaculty of Foreign Language, Daito Bunka University
1-9-1 Takashimadaita, Itabashi, Tokyo 175-8571, Japan

^bFaculty of Technology, Miyazaki University
Gakuen Kihanadai Nishi, Miyazaki, 889-2192, Japan

^cGraduate School of Pharmaceutical Sciences, University of Tokushima
1-78 Sho-machi, Tokushima 770-8505, Japan

^dResearch Institute of Computational Science, National Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

*e-mail: u.nagashima@aist.go.jp

(Received: August 31, 2006; Accepted for publication: October 20, 2006; Published on Web: December 1, 2006)

Parameters as an index to efficiently assess the pollution level of the upper, middle, and lower streams of the Tamagawa (Tokyo, Japan) based on measured water quality were determined by using multivariate analysis, principal component analysis (PCA), and cluster analysis (CA) for data measured from 1994 to 2002. Missing data during 2000-2002 were estimated using a perceptron type neural network and arithmetic progression. The combination of scores for the first and second principal components obtained by PCA enabled classification of the upper, middle, and lower streams of the Tamagawa. The CA results corresponded well with the PCA results.

Based on the score of the first principal component determined here, contributions to the water pollution of the middle and lower streams should be decreased to improve the water quality of the Tamagawa.

Keywords: Tamagawa, Principal Component Analysis, Cluster Analysis, Water Contamination, Chemometrics

1 Introduction

Rivers flowing through megacities are changing rapidly due to changes in human lifestyles. For example, massive water-intake from the rivers is now needed, and treated sewage water is dumped into the rivers. Water pollution in rivers flowing close to megacities significantly affects the life of every person living or working near the river. Water quality of such rivers has been extensively researched by using chemometric techniques. For example, by using cluster analysis (CA), principal component analysis (PCA), discriminant analysis (DA), and factor analysis (FA), Kowalkowski *et al.* [1] found that certain locations of the Braba River in Poland are influ-

enced by municipal contamination and/or agriculture. By using three multivariate techniques (FA, PCA, and DA), Alberto *et al.* [2] reported spatial and temporal changes in the Suquia River in Argentina. Santos-Roman *et al.* [3] used the same multivariate techniques (FA, CA, and DA) to develop equations that could predict certain water quality conditions for unmonitored watersheds in Puerto Rico. Using PCA, St-Hilaire *et al.* [4] confirmed the importance of nutrients as variables explaining a significant portion of variance for both freshwater and estuarine stations, and using CA, they concluded that for the Richibucto River in Canada, high nutrient concentrations associated with peat harvesting and municipal effluent are likely greater causes of concern than farming.

In this study, the purpose was mainly to develop a means to decrease the pollution of the Tamagawa in Tokyo, Japan. By using principal component analysis (PCA) and cluster analysis (CA) of water quality data measured yearly from 1994 to 2002, we determined parameters that can be used as an index to efficiently determine the pollution level at different locations along the Tamagawa.

2 Tamagawa

The main streams of the Tamagawa (Figure 1) start at the Okutamako Lake and then flow into Tokyo Bay through the western suburbs of Tokyo metropolitan area. The Tamagawa is 138km long, and its catchment covers an area of 1240km². Approximately 4,240,000 people live and/or work along this river. This river is a typical urban river in Japan; water pollution is caused by the intake of clean water from the upper stream and by the drastically yearly increase in inflow of treated sewage water in the middle stream, reflecting the high economic activity and urbanization. To decrease water pollution, an index needs to be determined that can describe the Tamagawa and can be used to evaluate the water quality based on measured water quality data.

3 Data set and Statistical Procedures

Water quality data for 17 observations points along the Tamagawa and its branches (Figure 1) for February for FY 1994-2002 measured by the Tokyo Metropolitan Government Bureau of Environment were obtained from the website of the Tokyo Metropolitan Government [5]. In this study, only 12 of the 90 possible indexes of water quality were used in our analysis (Table 1) because these 12 indexes are quantitative chemical parameters. From 1994 to 1999, all 12 indexes were reported, whereas from 2000 to 2002, some data were missing. At observation points No. 12 and 14, water quality was not measured after 2001. Cl⁻ was not monitored at 5 observation points (No.1, 3, 5, 12 and 14) after 2000, and at 7 observation points (No.2, 4, 7-9, 11 and 13) after 2001. COND was not monitored at 8 observation points (No.2, 4, 7-9, 11, 13 and 14) after 2001 and at No.12 after 2002.

Due to this missing data, analysis could not be done using conventional methods such as PCA. Therefore, in this study, the missing data were estimated using a perceptron type neural network and arithmetic progression [6]. Due to differences in scale of data, data used were then standardized for comparison. All values had compatible units from a distribution with a mean of 0 and a standard deviation of 1.

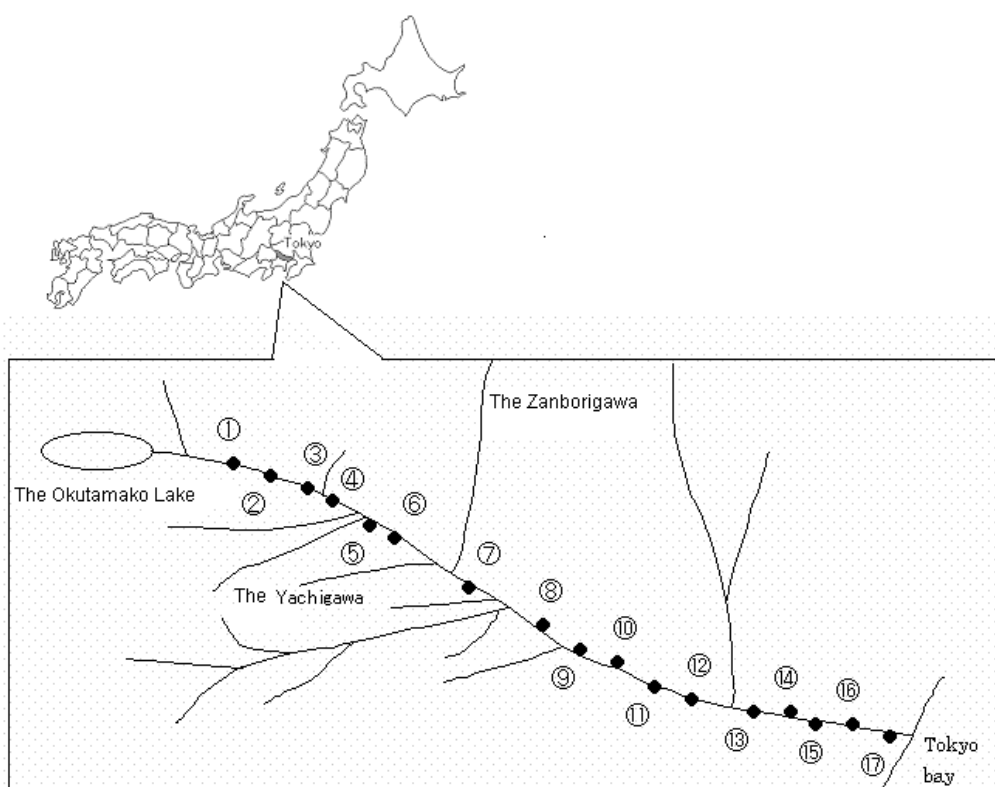


Figure 1. Schematic map of observation points along the Tamagawa, Tokyo, Japan.

Table 1. Chemical indexes used in water analysis of the Tamagawa

Chemical index (abbreviations)	Explanation
Activity of hydrogen ions (pH)	Water acidity or alkalinity. pH is high in limestone areas and/or areas close to the sea.
Dissolved Oxygen (DO)	Amount of dissolved oxygen that is freely available in water to sustain fish and other aquatic organisms. Lower DO level indicates higher level of water pollution.
Biochemical Oxygen Demand (BOD)	Total amount of oxygen consumed in the biological processes that break down organic matter in water. Higher BOD level indicates higher level of water pollution.
Chemical Oxygen Demand (COD)	Mass concentration of oxygen consumed by the chemical breakdown of organic and inorganic matter. Higher COD level indicates higher level of water pollution.
Total Nitrogen (T-N)	Total amount of nitrogen compounds contained in the water. Can be divided into inorganic and organic matter groups, as well as into dissolved matter and particulate matter groups.
Total Phosphorus (T-P)	Total amount of phosphorus compounds contained in the water. Can be divided into inorganic and organic matter groups, as well as into dissolved matter and particulate matter groups.
Chloride ion condensation (Cl ⁻)	Total amount of chloride ions. Index of pollution caused by human activities.
Ammonium Nitrogen (NH ₄ -N)	Total amount of ammonium ions mainly from pollution from human and animal waste. High NH ₄ -N levels increase nitrification, and lower the DO level.
Nitrite Nitrogen (NO ₂ -N)	Total amount of nitrite ion mainly from pollution from agricultural fertilizers. Chemicals are formed in the decomposition of waste materials, such as manure or sewage.
Nitrate Nitrogen (NO ₃ -N)	Total amount of nitrate ions mainly from pollution from agricultural fertilizers. Chemicals are formed in the decomposition of waste materials, such as manure or sewage.
Phosphate Phosphorus (PO ₄ -P)	Total amount of nitrate ions mainly from pollution from agricultural fertilizers, detergents, etc.
Conductivity (COND)	A measure of the ability of water to carry an electrical current. An increase in ion concentration causes an increase in COND.

Principal component analysis (PCA) was then applied to the combined data set to determine parameters that can be used as indexes to efficiently describe the pollution level of the Tamagawa. The aim of PCA is to find and interpret hidden complex and causally determined relationships within a data set.

Cluster analysis (CA) was then used to confirm the reliability of parameters obtained using PCA. CA is a classification method that is used to group individuals or variables. CA results were presented as dendrograms obtained using normalized Euclidean distances and the Ward's method.

All calculations in this study were performed by applying SPSS 11.0 software running on a Windows XP platform.

4 Results and Discussion

4.1 Principal component analysis (PCA)

Table 2 lists the eigenvalues and contributions of the principal components. Eigenvalues of the first, second, and third principal components were 7.34, 2.11, and 0.92, respectively, and the respective contributions were 61.4%, 17.5% and 7.7%. We discuss only the results for the first and second principal components whose eigenvalues were greater than 1. The combined contribution of the first and second principal components was greater than 10%, and the cumulative contribution rate was 78.9%.

Table 3 lists the coefficients of the indexes in the first and second principal components. In the first principal component, signs of both pH and DO are negative, and absolute values of Cl⁻ and COND are low, less than 0.3. Therefore, except Cl⁻ and COND, all indexes are important for the first principal component. The first princi-

pal component was complicated and consisted of a linear combination of ten chemical indexes. In the second principal component, Cl^- and COND are high and significant. This suggests that the parameter of combination

of indexes such as the first principal component is effective to analyze the degree of water contamination in the Tamagawa.

Table 2. Eigenvalues and contribution (%) of principal components.

Component	Eigenvalue	Contribution	Cumulative contribution
1	7.37	61.39	61.39
2	2.11	17.55	78.93
3	0.92	7.66	86.59
4	0.57	4.79	91.38
5	0.40	3.33	94.71
6	0.22	1.81	96.52
7	0.19	1.57	98.09
8	0.11	0.89	98.98
9	0.05	0.42	99.40
10	0.04	0.32	99.72
11	0.02	0.19	99.91
12	0.01	0.09	100.00

Table 3. Coefficients in the first and second principal components

Index	First component	Second component
pH	-0.488	0.161
DO	-0.856	-0.167
BOD	0.787	0.093
COD	0.973	0.031
T-N	0.955	-0.027
T-P	0.952	-0.199
Cl^-	0.211	0.952
$\text{NH}_4\text{-N}$	0.705	0.247
$\text{NO}_2\text{-N}$	0.893	-0.106
$\text{NO}_3\text{-N}$	0.875	-0.227
$\text{PO}_4\text{-P}$	0.924	-0.273
COND	0.229	0.948

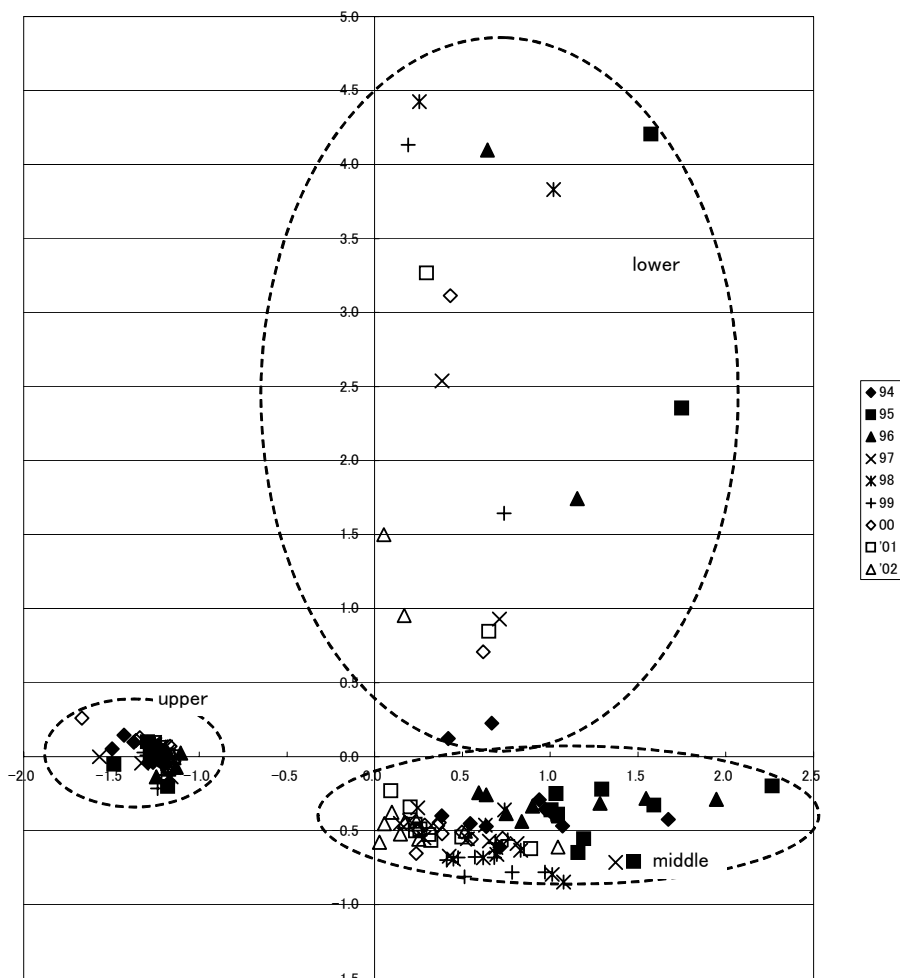


Figure 2. K-L plot of principal component scores. X- and Y-axes are the first and second principal components, respectively.

Figure 2 shows a K-L plot of the principle component scores of 153 data points, where the X- and Y-axes are the first and second principal components, respectively. The data points near the point (-1.5, 0.0) correspond to the upper stream of the Tamagawa, those in the region (X>0, Y<0) correspond to the middle stream, and those in the widespread region (X>0, Y>0) correspond to the lower stream. The first principal component classifies the data into two groups on the K-L plot: the upper stream and the others, namely combined middle/lower stream. This indicates that the first principal component can be used to classify the upper stream and the combined middle/lower stream, but can not distinguish the middle and lower streams.

The observation points corresponding to the lower stream have high absolute values of the second principal component, whereas those of the upper and middle streams have either low or negative absolute values. In contrast, the second principal component can be used to roughly classify the data into the combined upper/middle stream and the lower stream. Combination of the scores for the first and second principal components therefore enables classification of the data into the upper, middle, and lower streams of the Tamagawa.

Figure 3 shows the change in the average score of the first principal component for the entire river (all three

streams combined), the upper stream, middle stream, and lower stream during FY 1994-2002. A smaller score for this component indicates higher water quality.

The equation for the regression line for the score was $y = 0.0107x - 1.3078$ ($R^2 = 0.2206$) for the upper stream, $y = -0.1097x - 1.2403$ ($R^2 = 0.7214$) for the middle stream, and $y = -0.1022x - 1.1616$ ($R^2 = 0.4227$) for the lower stream, where R is the correlation coefficient. The slope of the regression line of the upper stream is almost zero, indicating that the water quality of the upper stream of the Tamagawa has remained relatively clean and constant from 1994 to 2002. The slope of the regression line of the middle stream is almost the same and that of the lower stream is negative, suggesting an increase in water quality during this observation period. In addition, contamination of all three streams increased significantly in 1995 and slightly in 1998. The slopes of the regression lines are the same for the middle and lower streams, indicating that improvement in water quality in these two streams significantly affects the water quality of the entire river.

Figure 4 shows the change in the score of the first principal component for three groups of fiscal years (1994-1996, 1997-1999, and 2000-2002) and the average at each observation point. Again, a smaller score indicates higher water quality.

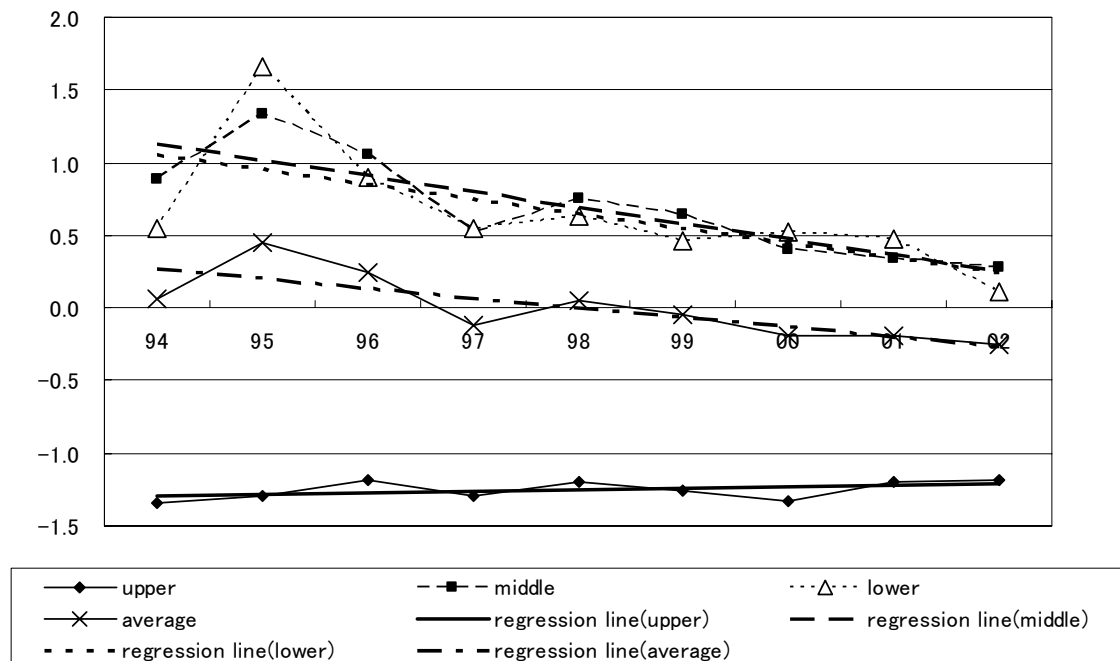


Figure 3. Change in average score of the first principal component for the entire stream (all three streams), upper stream, middle stream, and lower stream from 1994 to 2002.

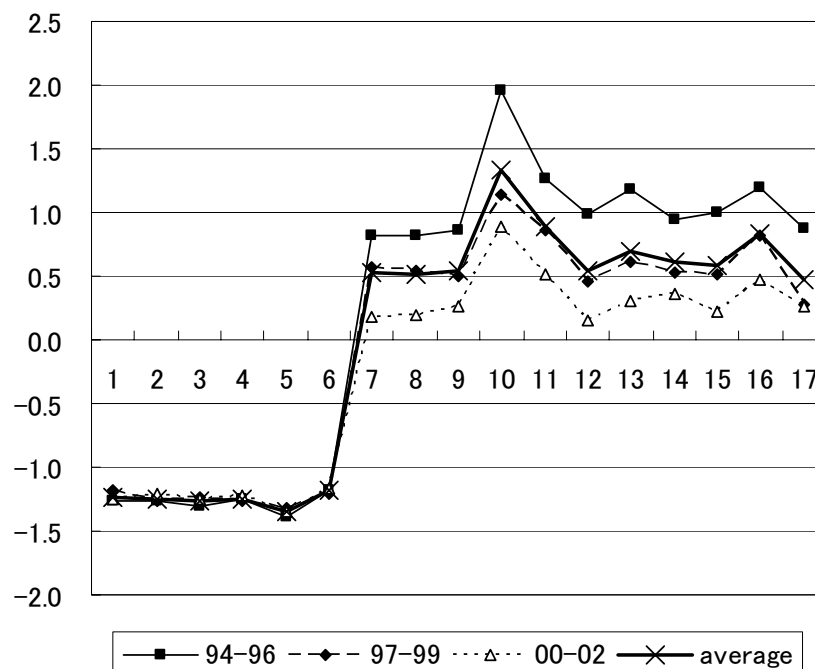


Figure 4. Change in the score of the first principal component for the entire observation period (1994-2002) and for three groups of FY (1994-1996, 1997-1999, and 2000-2002) at every observation point.

In the upper stream (No.1-6), the scores are negative. In contrast, the scores for the middle and lower streams (No.7-17) are positive, indicating a significant increase in contamination of these two streams, corresponding to the results in Figure 3.

The score for first principal component changes drastically from negative at No.6 to positive at No.7, suggesting a strong influx of pollutants between these two locations in the river. To determine the cause of this contamination, we investigated relationships between the water quality of the main stream and the tributaries (the Yachigawa and the Zanborigawa; Figure 1) that flow into this main stream between these two observation points. The water quality of the Yachigawa River is more contaminated than the main stream at No. 6 and 7 in terms of BOD, COD, $\text{NH}_4\text{-N}$, $\text{NO}_2\text{-N}$, and $\text{NO}_3\text{-N}$, whereas the water quality of the Zanborigawa River is more contaminated in terms of T-P, $\text{PO}_4\text{-P}$. All these contaminants strongly affected the water quality at No.7.

In addition, the change in water quality between No.9 and No.10 was also significant. However, we were not able to determine the source of pollutants. The score of the first principal component sharply decreased between No.10 and No.12, indicating a mechanism of water purification. A weir and some holms exist between No.10 and 11, and between No.11 and 12, possibly purifying

the water via back water. The purification mechanisms of weirs and holms remain unclear. As evident in Figure 3 as well, improvement in water quality in the middle and lower streams significantly affected the water quality for the entire river.

4.2 Cluster analysis (CA)

Figure 5 shows a dendrogram of the location pattern based on the CA of measured data from 1994 to 2002. All of the observation locations could be generally grouped into five main clusters. First, cluster V was separated from the other clusters. Cluster V is a group of data from the lower stream from 1995 to 2001. In this period, the lower stream was highly polluted. Second, cluster I was separated from the other clusters. Cluster I is a group of data from the upper stream from 1994 to 2002, and is the clearest group of data. Third, cluster IV was separated from the other clusters. Cluster IV is a group of data from the polluted middle stream from 1994 to 1996. Among the remaining data, cluster II is a group of data from the lower stream that are not included in cluster V. Cluster II is a group in which water quality has improved during the entire observation period, as evidenced in Figure 3. In the middle stream, the water quality of cluster III was higher than that of cluster IV.

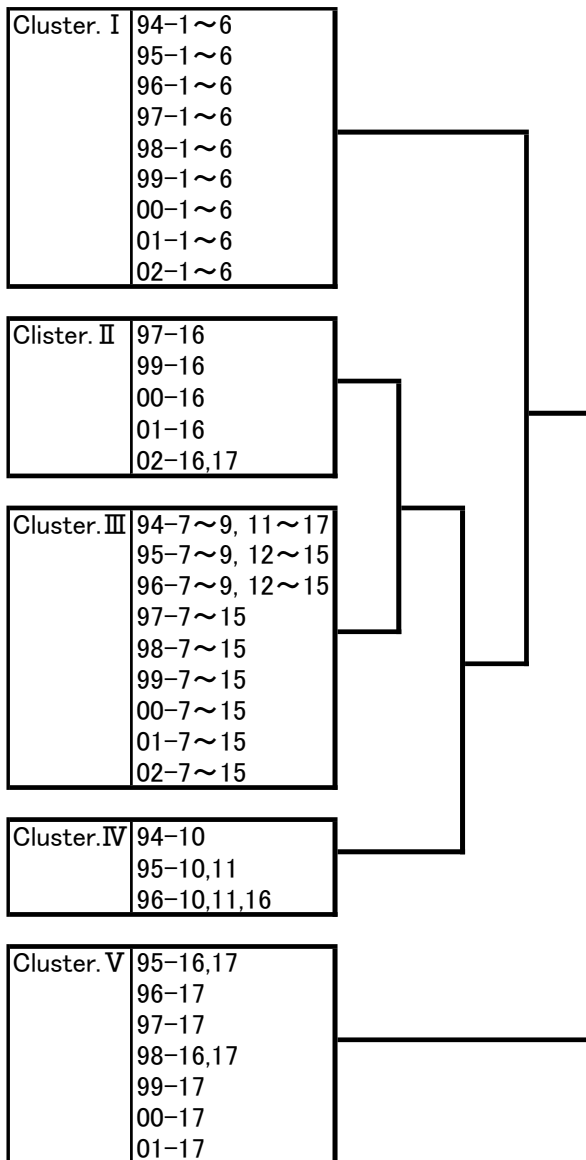


Figure 5. Dendrogram of the location pattern for the entire observation period from 1994 to 2002.

Based on the dendrogram, the observation points can be roughly assigned to three groups: the upper stream, which is clean; the lower stream, which is highly polluted; and the middle stream, which is intermediate in pollution level between the upper and lower streams. This grouping corresponds well with that in the K-L plot from the PCA (Figure 2).

Figure 6 shows a dendrogram of the 12 indexes based on the CA analysis. According to this dendrogram, all 12 indexes can generally be grouped into three main clusters. First, pH and DO were separated from the other indexes, and correspond well with the negative coefficients of pH and DO in the first principal component. Second, Cl⁻ and COND were separated from the other indexes. These two indexes had high values in the lower stream due to the influence of seawater and/or severe pollution.

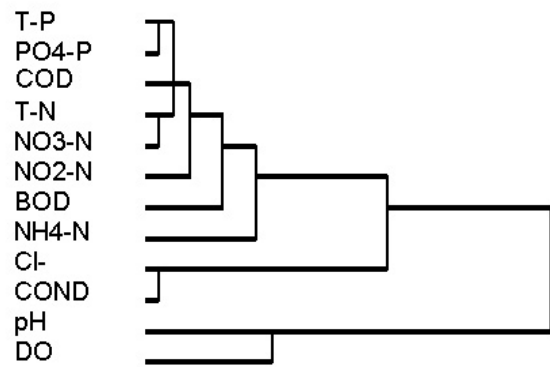


Figure 6. Dendrogram of the 12 chemical indexes

Such high values correspond to the high coefficients of Cl⁻ and COND in the second principal component. The remaining indexes are included in the other cluster. The CA results correspond well to the PCA results.

5 Conclusions

The aim in this study was to obtain a compound parameter to assess the pollution levels along different streams of the Tamagawa in Tokyo, Japan, by using water quality data measured during 1994-2002. Certain data were missing for 2000-2002, and therefore estimated here using a perceptron type neural network and arithmetic progression.

Principal component analysis (PCA) and cluster analysis (CA) were then applied to the combined data set to obtain a parameter that can efficiently assess the water quality of the Tamagawa. The first principal component was complicated and consisted of a linear combination of ten chemical indexes. The second principal component was a linear combination of two indexes. Combination of the first and second principal components enabled the classification of the upper, middle, and lower streams of the Tamagawa with respect to pollution level. Changes in the average score of the first principal component for the entire stream (i.e., all three streams combined), upper stream, middle stream, and lower stream suggested that the water quality of the upper stream did not change during the observation period and that of the middle and lower streams slightly improved. We confirmed the change in the score of the first principal component for the entire observation period (1994-2002) and for the three groups of FY (1994-1996, 1997-1999, and 2000-2002) for each observation point. In the upper stream (No.1-6), the first principal component

scores are negative. In contrast, the scores for the combined middle/lower streams (No.7-17) are positive, indicating an increase in contamination of the combined middle/lower streams. The first principal component score changes drastically from negative at No.6 to positive at No.7, indicating a major source of pollution entering the stream between these two locations. The water quality of both branches that flow into the Tamagawa between No.6 and 7 is more contaminated than the Tamagawa itself. The first principal component score sharply decreases between No.10 and 12, indicating a source of purification, such as a weir or holms.

The CA results corresponded well with the PCA results. The dendrogram of the location pattern shows that all the monitoring locations can be generally grouped into five main clusters. The dendrogram of the 12 chemical indexes shows that all the indexes can be generally grouped into three main clusters.

Based on these results, water contamination in the middle and the lower streams should be reduced to improve the overall water quality of the Tamagawa.

We thank Professor H. Chuman of Tokushima University for numerous stimulating discussions.

References

- [1] Kowalkowski, T., Zbytniewski, R., Szpejna, J., Buszewski, B., Application of chemometrics in river water classification, *Water Res.*, **49**, 744-752 (2006).
- [2] Alberto, W. D., Del Pilar, D. M., Valeria, A. M., Fabiana, P. S., Cecilia, H. A., De Los Angeles, B. M., Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: suquia river basin (Cordoba-Argentina), *Water Res.*, **35**, 2881-2894 (2001).
- [3] Santos-Roman, M. D., Warner, S. G., Scatena, F., Multivariate analysis of water quality and physical characteristics of selected watershed in Puerto Rico, *Journal of the American Water Resources Association*, 829-839 (2003).
- [4] St-Hilaire, A., Brun, G., Courtenay, C. S., Ouarda, B. M. J. T., Multivariate analysis of water quality in the Richibucto drainage basin (New Brunswick, Canada), *Journal of the American Water Resources Association*, 691-703 (2004).
- [5] Tokyo Metropolitan Government Bureau of Environment., 1994-2002. Kokyoyo suiiki suishitsu sokutei kekka ("Reports of quality of water for supply"). Electronic publication, <http://www2.kankyo.metro.tokyo.jp/kansi/mizu/sokutei/sokuteikekka/kokyou.htm>. (in Japanese).
- [6] Kambe, J., Yan, Y., Nagashima, U., Aoyama, T., *J. Comput. Chem. Jpn.*, **5**, 201-212 (2006), (in Japanese).