Factor and Cluster Analyses of Water Quality in Kansas River

○ Rim, Chang Soo*, Jo, Kwan Hyung*

1. Introduction

Variations of ambient surface-water quality in a river basin are influenced by factors such as geologic and climatic characteristics and human activities in the river basin. The causal relationship between factors and the observed values of chemical, physical, and biological parameters in the river basin cannot be established directly. Therefore, surface-water quality is usually defined based on observations of chemical, physical, and biological parameters in the river basin. From the systems point of view, the river is a system and the observed water quality parameters are the system output. However, the input to the system is unknown. Therefore, a causal model cannot be established to relate the measured water-quality time series to the unknown input.

In the last decade, U.S. national and state fixed-station stream quality monitoring networks have been established. A lot of surface-water quality data have been collected and stored by EPA. Analysis of the lower Kansas River data was conducted by the USGS researcher (Jordan and Stamer, 1991), as part of the National Water Quality Assessment Program (NAWQA) (Hirsch et al., 1988).

The purpose of this study is to make some exploratory investigation on the application of multivariate statistical analysis to selected stream-quality data in the lower Kansas River basin. In this study, SPSS statistical software was used for factor analysis and cluster analysis. Factor analysis is used to postulate a model in which each observed variable is dependent on a set of unobservable or latent variables. The number of latent variables or factors is much less than the number of variables. Therefore, the dimension of the problem is reduced and perhaps, these factors can be interpreted in a meaningful way.

In addition, cluster analysis is also performed to determine the similarity and dissimilarity of water quality characteristics among the sampling stations. This information is valuable in evaluating the cost-effectiveness of the existing monitoring network.

2. Data

The data summary retrieved from EPA indicates that Kansas Department of Health and Environment has compiled chemical water quality data since September 25, 1948. As of 1989,

* Department of Civil and Environmental Engineering, Chungwoon University

there are 134 fixed stations with incomplete data distributed on major rivers and their tributaries. The sampling periods, sampling frequencies and sampling parameters, however, vary from station to station and also vary from year to year. The sampling frequencies are no higher than once a month. Among the 134 stations, only Station 280 on the Ninnescah River near Belle Plaine of the Arkansas River basin has chemical quality data that began in 1948. Most of the other stations began collecting data in 1967 or later.

A preliminary investigation of the data shows that there are gaps, and/or missing data in most of the records. Therefore, only data collected at nine stations (Nos. 257, 258, 259, 260, 262, 101, 237, 238 and 239) in the lower Kansas River basin as shown in Fig. 1 were selected for multivariate statistical analysis. These records have relatively less missing monthly values than those at other stations.

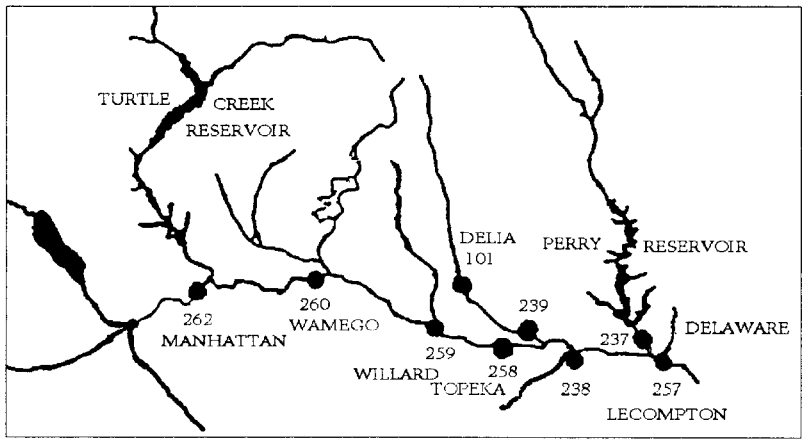

Fig. 1. Water Quality Sampling Stations of Study Area.

Monthly values from May 1977 through September 1981 for nine parameter were used in the analysis. They are biochemical oxygen demand (BOD, mg/*l*), chemical oxygen demand (COD, mg/*l*), specific conductance (SC, micromhos/cm), total coliform (TC, mpn/100*l*), turbidity (TURBI, mgSio_2/*l*), dissolved oxygen (DO, mg/*l*), ammonium (NHN, mg/*l*), phosphorus (P, mg/*l*), and stream flow (cms). It should be noted that records at Stations 238, 259, and 262 contain no stream flow data. No attempts were made to estimate stream flow at these stations from available records at nearby stations.

Stations 259, 262 and 238 have no streamflow measurements. Therefore, to consider the influence of streamflows on water-quality characteristics, only data collected at 6 stations were analyzed. Missing values of water-quality parameters in ith month of jth year (MV_{ij}) at each station were estimated based on the following equation.

$$MV_{ij} = \frac{\Sigma(Q_{ij} \cdot V_{ij})}{\Sigma Q_{ij}} \qquad (1)$$

where i is the month of the year during the study period at each station; j is the year during

the study period at each station; Q_{ij} is streamflow rate of ith month of jth year during the study period at each station; and V_{ij} is value of water quality variable measured in the ith month of jth year during the study period at each station. Therefore, $\sum(Q_{ij} \cdot V_{ij})$ is the summation of multiplication of streamflow and water quality values during the study period at each station. $\sum(Q_{ij})$ is the summation of streamflow during the study period at each station. However, because Stations 101, 238, and 239 have many missing values, using Eq. (1) for estimating the missing data may distort the true statistical nature of the data set.

The means, standard deviations and coefficients of skewness of the original data were computed. The results show that the data sets all have relatively small or large values of positive coefficient of skewness. Therefore, the original data were log-transformed to achieve approximately normal distribution.

The descriptive statistics of log-transformed data at nine stations shows that most variables have approximately normal distribution at each station. To avoid the occurrence of nonpositive number of log-transformed values or infinity, a positive number was added to the original value prior to transformation. In addition, DO value was inverted to show that increasing in the value of DO indicates degradation of water quality. Furthermore, the inverted dissolved oxygen(DO^*) value multiplied by 100 was used in the analysis. Thus, for all 8 water-quality parameters, increasing in the value of each parameter indicates degradation of water quality.

3. Methods of Analysis

3.1 Analysis of Variance (ANOVA)

The analysis of variance was performed. In the case of BOD, F ratio = 1.91 (F probability=0.109) < F ratio = 2.37 [(d.f. 4, 259), F probability = 0.05], therefore, null hypothesis (H0) is accepted. This means that there are no significant differences in the BOD mean values at 5 stations. Similarly, the mean values of COD, TURBI and P show no significant differences. On the other hand, the tests show that significant differences exist in the means of SC, TC, DO^* and NHN, especially TC and SC.

3.2 Factor Analysis

Factor analysis is one of a class of methods of multivariate statistical analysis whose primary purpose is data reduction and summarization (Anderson, 1986). Its objectives are to analyze the interrelationships among a large number of variables and to explain these variables in terms of their common underlying dimensions known as factors.

3.2.1 Correlation Analysis of Water Quality Variables

Most variables showed very high, high, low and almost ignorable correlation at each station. Most variables do not show significant correlation at 5 different stations. It is

apparent that station 259 and station 262 showed much higher correlations of water quality variables than those of other stations. However, considering the whole correlation of each variable of each station, it could be understood that there are some identical correlation between variables at each station, but not obviously. It is very interesting to see that SC has inverse correlation with all other variables at all 5 stations. Only correlation between TURBI and COD shows high and very high values at 5 stations. Correlation between TURBI and SC, TURBI and P show high and very high values at 5 stations except the correlation between TURBI and SC at station 258. On the other hand, correlation between BOD and SC, BOD and TC showed almost ignorable correlation between variables at all stations. As a result, it can be concluded that correlations of each variable are not significant through whole stations; however, it seemed that there are some identical water quality characteristics through whole stations.

3.2.2 Estimation of Factor Loadings and Factor Scores

Three factors are obtained at each station to increase the percentage of variance explained by factors, and to make it easier to interpret factors at each station. Table 1 shows the factor loadings of each factor obtained from corresponding eigenvalues and eigenvectors at Station 257. Double line was used to indicate the variables which are highly loaded in each factor.

Table 1. Factor Loadings and Communality at Station 257

	FACTOR 1	FACTOR 2	FACTOR 3	COMMUNALITY
TURBI	0.892	0.180	-0.053	0.832
P	0.795	-0.031	-0.193	0.671
SC	-0.755	-0.017	-0.138	0.590
COD	0.562	0.183	0.195	0.388
NHN	0.300	-0.840	0.123	0.811
DO*	0.271	0.834	-0.022	0.770
BOD	0.300	0.727	0.159	0.664
TC	0.016	-0.011	0.963	0.928

The information obtained from factor loadings provides some meaningful interpretation. However, it must be remembered that there is no clear-cut method to interpret each factor, which means that it is very subjective depending on each investigator's opinion. At Factor 1, apparently, TURBI, COD, P are highly loaded at all 5 stations, which means that they have some similar water-quality characteristics. Turbidity has some relationship with erosion and land use. Phosphorus is related to fertilizer use. Therefore, Factor 1 perhaps represents the effects of land use and non-point source of pollution. On the other hand, at some stations, DO*, BOD and NHN are highly loaded for factor 2. Factor 2 can be interpreted as nitrification effects. However, it may not be significant because changes in NHN means of 5 stations are not significant. Factor 3 may be inferred as point-source pollution effects.

3.2.3 Factor Scores

Considering only 5 stations along the main stream (257, 258, 259, 260, 262), the changes in scores of factor 1 at each station do not show clearly any trend in water quality. The factor scores at each of the nine stations are obtained for cluster analysis. The mean value of factor scores for each factor should be zero.

3.3 Cluster Analysis

The single-linkage agglomerating hierarchical method (eq. 2) is used in this study for clustering the 9 fixed stations. The proximity of individuals is usually explained as a distance such as the Euclidean distance. This is the distance summed by the square of difference between variables, at two stations 1 and 2 (d_{12}) as

$$d_{12} = [\ \sum_{i}(X_{i1} - X_{i2})^2\ ]^{0.5} \qquad (2)$$

where, X_{i1} and X_{i2} represent the stations 1 and 2, respectively. In this study, the median value of the i th factor score is used.

Table 2 shows the cluster membership of using 3 factors. Stations of main streams 262, 260, 259, 258 and 257 have memberships in cluster 1 when they are clustered into two or three clusters. However, Station 260 was not clustered with other Stations 262, 259, 258 and 257 when four and five clusters are used. Probably, that's because station 260 has lower factor scores in each factors compared with other stations (262, 259, 258, 257) of the mainstream.

Table 2. Cluster Membership Based on Median Scores of 3 Factors

	Number of Clusters			
	5	4	3	2
STA 257	1	1	1	1
STA 258	1	1	1	1
STA 259	1	1	1	1
STA 260	2	2	1	1
STA 262	1	1	1	1
STA 101	3	3	2	2
STA 237	4	3	2	2
STA 238	1	1	1	1
STA 239	5	4	3	2

4. Relationship Between Streamflow and Water quality Characteristics

The relationship between streamflow and water-quality characteristics was examined. Stations 257, 258, 260 and 239 show high, low and almost ignorable correlation between stream flow and water-quality variables. On the other hand, Stations 101 and 237 showed

low or almost ignorable correlation between water flow and water quality variables. It should be noted that Station 237 is located below Perry Reservoir. Therefore, its water quality characteristics are highly influenced by reservoir operation. Turbidity has high and low correlation with water flow at all stations, except stations 101 and 237 which have almost ignorable correlation with streamflow. BOD and TC showed the almost negligible correlation at all stations. SC showed the inverse correlation with streamflow at all stations, especially, it showed the highest inverse correlation at Station 257. DO* showed almost ignorable correlation with streamflow at all stations except Station 257. Considering the fact that non-point source pollution effects are positively correlated with streamflow, it is expected that those water-quality variables show higher correlation than those of other variables with streamflow.

5. Conclusions

The results of multivariate statistical analysis of selected data for the lower Kansas River basin lead to the following conclusions:

(1) There are significant correlations among the eight water-quality variables. The variability of the median values of these variables at nine Stations is not substantial.
(2) Three factors can be identified to correlate water-quality variables. They are : land use and non-point source of pollutions, nitrification, and point source of pollutions. The water quality characteristics may be affected by one or more of these factors.
(3) The cluster analysis shows that Stations 257, 258, 259, 260, 262 and 238 probably have similar water-quality characteristics while Stations 101 and 237 belong to another cluster and Station 239 would stand alone.

6. References

1. Anderson, T.W. (1986). An introduction to multivariate statistical analysis, NewYork, West Publishing Company.
2. Barlett, M.S. (1937). "Properties of sufficiency and statistical tests." Proceedings of the Royal Society of London (A), Vol. 160. pp. 268-282.
3. Barlett, M.S. (1938). "Further aspects of the theory of multiple regression." Proceedings of the Cambridge Philosophical Society, Vol. 34, pp. 33-40.
4. Hirsch, R.M., Alley, W.M. and Wilber, W.G. (1988). "Concepts for a national water-quality assessment program." U.S., Geol. Surv., Circ. No. 1021.
5. Johnson, R.A. and Wichern, D.W. (1988). Applied multivariate statistical analysis, 2nd ed., Prentice Hall, Inc., Englewood Cliffs, New Jersey.
6. Jordan, P.R. and Stamer, J.K. (1991). "Surface water-quality assessment of the lower Kansas River basin, Kansas and Nebraska: Analysis of available data through 1986. U.S. Geological Survey, Water-Supply Paper 2352-B.
7. Krzanowski, W.J. (1988). Principles of multivariate analysis. Oxford Science Publications Inc.